Heuristic Model to Improve Feature Selection Based on Machine Learning in Data Mining

Jahin Majumdar
Department of CSE, ASET
Amity University
Noida, India
jahin07@gmail.com

Anwesha Mal
Department of CSE, ASET
Amity University
Noida, India
anweshamal@gmail.com

Shruti Gupta
Department of CSE, ASET
Amity University
Noida, India
sgupta6@amity.edu

***Abstract*— Data Mining and Machine Learning is one of the most popular research areas in computer science that is relevant in today's world of unfathomable data. To keep up with the rising size of data, there arises a need to quickly extract knowledge from data sources to aid data analysis research and improve industry and market needs. Primary Data Mining algorithms like k-means, Apriori, PageRank etc. are used today, but Machine Learning techniques can enhance the same by learning from the complex patterns. This paper focuses on the various existing approaches where Machine Learning algorithms have been used to improve data classification and pattern recognition in Data Mining especially for Feature Selection. It compares and contrasts the existing techniques and finds out the best one among them. Further, the paper proposes a heuristic approach to theoretically overcome most of the limitations in existing algorithms.**

Keywords—Data Mining; Machine Learning; Fuzzy; Knowledge Discovery

I. INTRODUCTION

Machine Learning is defined as the ability of computers to learn from without being explicitly programmed. According to Tom Mitchell of Carnegie Mellon University [1], "A computer program is said to learn from Experience 'E' with respect to a Task 'T' and some performance measure 'P', if its performance on T as measured by P improves with E." Machine Learning algorithms can learn from a predefined dataset and predict the results of a test dataset. Machine Learning can be divided into two broad categories, namely a) Supervised Learning b) Unsupervised Learning and c) Reinforcement Learning [2]. In Supervised Learning, a training dataset is provided which consists of an example pair. The example pair consists of the input object and the expected output value. The algorithm can analyze the given data and produce a function to predict or map the outputs of the remaining data. In Unsupervised Learning, the computer is expected to find the pattern on its own without the help of labels in the training dataset. The goal of Unsupervised Learning is to discover hidden patterns in a large pool of data. Reinforcement Learning enables the system to perform its task in a dynamic environment without teaching it whether it has come closer to achieving its goal or not.

Data Mining is defined as the process of finding patterns in a large data set and is a part of Knowledge Discovery in Databases (KDD) [3]. In fact, Data Mining is the fourth step in the KDD process. The ultimate aim of Data Mining is to convert a raw set of data into a meaningful structure for later use. According to [3], Data Mining involves six main classes of tasks namely, a) Anomaly Detection b) Associate Rule Learning c) Clustering d) Classification e) Regression and f) Summarization. Anomaly Detection is used to detect errors in a large pool of data. Association rule learning is another word for market basket analysis where it searches for patterns or relationships between variables. Clustering groups data according to various parameters. Classification labels new data according to a known structure. Regression uses a function to model the data with minimum error. Summarization presents a compact representation of a dataset.

Feature Selection is a problem where Machine Learning can be used in Data Mining. In Feature Selection, a subset of features is selected from a larger set of features, given that the subset is sufficient to describe the target process [4]. Feature Selection aims to obtain a feature space with the following properties a) dimensionality should be reduced b) sufficient information should be retained c) Effects arising due to noisy features should be eliminated which can improve the separability of feature selection and d) contrasting features in the same group. [5] Feature Selection aims to prevent selecting too many or too new necessary features. That the information content in the set of features is low can be understood if less number of features are selected. However, if irrelevant features are selected, the relevant data gets overshadowed by noise. Selecting the correct features is an important step in creating a predictive model. Unless the correct features are selected from beforehand or if feature selection is applied on all features, the overall accuracy of the process decreases. If more than two predictors are correlated to the predicted, then the coefficients in the regression model become unstable. Secondly, larger the set of predictors, higher is the probability of finding missing values in the data. Machine Learning can aid Feature Selection methods by improving the accuracy of the predictor function. Moreover, Machine Learning helps to select the features more accurately. Once the features are correctly selected, the accuracy of the algorithm will increase.

The rest of the paper is organized as follows: Section II discusses the existing work related to Machine Learning techniques implemented in Data Mining. Section III discusses the basic issues in Feature Selection with respect to a Heuristic Approach where a heuristic approach to Feature Selection is discussed. Section IV presents the heuristic model to resolve the issues in Feature Selection with appropriate techniques. Section V concludes the paper with potential applications.

978-1-4673-8203-8/16/$31.00

II. Related Work

A. Feature Selection Evaluation Technique for Learning

Piramuthu [6] evaluated a feature selection method and determined their effectiveness on preprocessing the input data to induce decision trees. The author evaluated these methods, especially the Sequential Forward Search (SFS) using various real world data. Both Probabilistic distance measures and inter-class distance measures were used. The evaluation technique used a cleaned Credit approval data from [8], whereby each data corresponded to a credit card application. The paper classified the results into Before Pruning (BP) and After Pruning (AP). In the training examples, a type of inter-class distance measure namely, the non-linear (Parzen and hyperspheric kernel) distance measure [9] scored an accuracy of 97.92% in BP while the Minkowski distance measure and the city block distance measure scored an accuracy of 97.68% in AP. In the case of Loan Default data, Minkowski and City Block scored an accuracy of 95.93% in BP while Euclidean and Mahalanobis scored an accuracy of 93.25% in AP. The Web traffic dataset shows Minkowski having the highest accuracy of 82.43% for BP and 81.38% for AP. The Tam and Kiang data [10] again had Minkowski as the technique with the highest accuracy of 94.64% for BP while Chebychev had an accuracy of 92.53% for AP. The paper showed that induced decision trees is sensitive to the input data and different scenarios. However, the paper failed to conclude the best distance measuring technique for real world scenarios.

B. Fuzzy Method in Machine Learning and Data Mining

In this paper, Hüllerman [7] used and reviewed fuzzy methods and their applications in Machine Learning and Data Mining. The paper described the typical applications of fuzzy set theory which included a) Fuzzy Cluster Analysis b) Learning fuzzy rule bases c) Fuzzy Decision Tree Induction d) Fuzzy Association Analysis e) Fuzzy Methods in case-based learning and f) Possibilistic Network. In Fuzzy Cluster Analysis, three main types of clustering algorithms are described namely, Mixture Modeling algorithm, Combinatorial algorithms and mode-seekers. An object may belong to multiple clusters with their membership degree expressed. The adaptation of rule-based model is described in Learning Fuzzy Rule bases. The author describes hybrid models that combine Fuzzy Set Theory (FST) and neural networks or evolutionary algorithms to optimize FST. The C4.5 [10] is a type of Decision tree induction algorithm where sets are partitioned recursively until homogenous subsets are obtained. In Fuzzy Association Analysis, the applications include market-basket analysis. Case-Based Learning, also called Instance-based Learning applies to very special cases of Machine Learning. Possibilistic Networks allow graphical representation but cannot solve uncertainty problems like imprecision or incompleteness. The paper further described potential areas where FST can contribute namely, a) Graduality b) Interpretability c) Robustness d) Representation of uncertainty e) Incorporating Background Knowledge and f) Generalized Aggregation Operators. The paper concludes by stating that FST can be an approach to solve Data Mining problems using Machine Learning. However, FST lacks in generalization performance and model accuracy.

C. Data Mining with Decision Trees and Decision Rules

Apté and Weiss [11] describes Decision Trees, tree and rule mining methodologies and how it can be used in Data Mining. The paper used various hypothetical examples to evaluate how decision trees can be used to improve Data Mining techniques. In Symbolic methods for classification modeling, the author took a hypothetical example for classifying pegs as a function of their length and diameter. The algorithm forms a tree like structure with the rules forming the branches. However, removing noise is a challenge. The error rate and the accuracy also needs to be checked with symbolic methods. Decision Trees can also be modeled in a top-down approach from the training dataset. The authors describe CART [12] which is a widely used binary tree modeling algorithm. The paper also describes the Pruning mechanism of CART to minimize error rate. C4.5 and CART is compared and C4.5 was found to have better predicting capabilities. Decision Rule Induction methods described attempt to correctly classify the given examples. It follows the Disjunctive Normal Form (DNF). Tree and Rule-based Regression is described to approximate the value if a continuous variable. Regression can be performed both by Tree Induction and Rule Induction as described in the paper. The paper mainly discusses about using regression modeling and decision trees to aid data mining techniques. However, consistent robustness across a large dataset prevents Symbolic Modeling from working well. The authors stress on using hybrid approaches to improve the modeling process while decision tress should be used for a sample with a single solution.

III. Issues in Feature Selection

Machine Learning is mainly used in Data Mining to solve the problem of Feature Selection which is the main content of this paper. As described before, Machine Learning aids the small number of features that are chosen from a larger dataset such that the subset is sufficient to describe the target process [4]. There are various approaches to Feature Selection problems. This paper focuses on a Heuristic Approach. Feature Selection has four basic issues namely, a) Determining an initial start point b) Organizing Search Space c) Evaluating alternate subsets of attributes and d) Finding a stop condition.

A. Determining an Initial Start Point

There are four main known methods for determining the start point namely [13], a) Forward Selection and Backward Elimination b) K-means Clustering c) Forward and Backward Stepwise Multiple Regression and d) Sequential Forward Search (SFS) and Sequential Backward Search (SBS).

Blum and Langley [13] describe that in Forward Selection, attributes are successively added after starting from nothing and in Backward Elimination, attributes are successively removed to arrive at the starting point. Mujtaba and Hussain [14] uses K-means clustering for partitioning data and finding cluster centers. These are used to find Radial Basis Function (RBF) nodes and find hidden layers in a network. In this process, total squared distance between the data points assigned to each cluster and the cluster center is minimized. The center is initialized to a random point in the space and each data point is assigned the cluster whose center is nearest to that data point. After assigning all the data points, the cluster center is assigned the mean (μ) of all the data points of that cluster. The data points are then

reassigned to the cluster with the nearest center and the process continues until there is convergence. Mujtaba and Hussain [14] also describes a Regression technique to determine center of the RBF nodes. An iterative approach is used to determine the centers of the nodes. The connection weights are simultaneously determined using a least square version. Linear Regression methods like Forward Regression and Stepwise regression successfully determine the hidden nodes. The last method is SFS and SBS [13] that obtains a chain of nested subset of features by addition or subtraction from the locally best or worst feature, as the case may be.

B. Organizing Search Space

Organizing Search Space is the second approach to implement Feature Selection. There are two main ways to organize the search space namely, a) Brute-Force approach and b) Greedy approach.

The Brute-Force approach is inefficient because for n attributes, there exists 2^n subsets. It has a high complexity and is impractical to apply. Therefore, a Greedy method is preferred. In the Greedy method, the current local subset of attributes is considered and that is considered to be the optimal solution for the whole set. Then, an iterative approach is followed to organize the complete search space.

C. Evaluating Alternate Subsets of Attributes

Induction Algorithms were used to evaluate the alternate subsets of attributes [15]. Induction Algorithms are types of algorithms that takes input specific instances and generates a model that generalizes beyond these instances. Decision Trees are also used to evaluate these alternate subsets. Two types of Decision Trees namely, a) Continuous and b) Discrete can be used in this case. In the case of supervised learning, a hypothetical space and a training dataset is required. There are two types of nodes present wherein each leaf node has a class label as determined by the training dataset and each internal node is a question of features which branches out according to the answers. In the C4.5 algorithm, a top-down greedy approach is followed. Decision trees are preferred because they are simple to use, understand, implement and computationally cheap.

Support Vector Machines (SVM) as described in Abraham et al. [16], can be used with associated learning algorithms that recognize patters and analyze data that are used for classification and regression analysis. It is essentially used to solve the Classification Problem. In the Classification Problem, two subsets are given in a large dataset and the aim is to generate a curve that can separate the subsets. SVMs finds and constructs a hyperplane that can separate the two subsets. In SVM, a linear combination is used to express a separating function $f(x)$ as a linear combination of kernels [16]:

$$f(x) = \sum_{x_j \in S} \alpha_j y_j K(x_j, x) + b$$

x_i denotes the training patters, $y_i \in \{+1, -1\}$ denotes the corresponding class labels and S the set of Support Vectors, b is the offset and α_i are the corresponding coefficients. SVM solves a quadratic optimization problem as understood from the training example. This solution uses numerical libraries to use optimization routines. However, optimizing these problems is computationally expensive. Moreover, SVM fails for inseparable data.

This is where Random Forests come into use. As described in Breiman [17], a Random Forest is a learning method that contains a combination of tree predictors such that the values of the random vector are dependent on each tree independently for the same distribution of all trees in the forest. The generalized error rate converges due to the dependence on the correlation between trees. Random Forests can be used effectively in prediction as the the overfitting problem is solved. The 'Randomness' in Random Forests improves the accuracy of classifiers and regressors. Random Forests, unlike other algorithms, keep the dataset constant. As such, Random Forests can be viewed as a Bayesian procedure. This makes Random Forests the preferred method to use for evaluating alternate subsets of attributes.

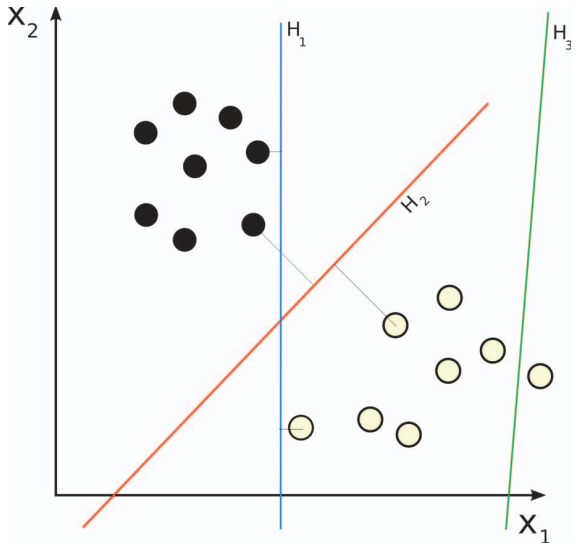

Figure 1. Three hyperplanes separating two data subsets with H_2 being the most effective

D. Finding a Stop Condition

This is the final issue in using a heuristic approach to Feature Selection. It denotes selection of the criterion to find a stop condition. In this case, attributes are no longer added after the alternate subsets improve the classification accuracy. The process is repeated until the accuracy significantly improves or until the end of the search space is reached. The halting criterion is to stop when each combination of values for those attributes successfully map onto a single class value. However, the data must be noise free for it to work. An alternative method is to use a relevancy score to order the set of features and determine the breakpoint using a system parameter.

IV. PROPOSED MODEL

The heuristic model proposed model consists of six major stages to improve Feature Selection to aid Data Mining techniques. It comprises of a) collecting the original dataset b) preprocessing the data c) selecting the feature subset using K-means clustering d) Applying SVM algorithm e) Checking for Halting condition and f) Displaying the resultant dataset. This heuristic model should theoretically overcome all the difficulties

faced in applying Feature Selection in Data Mining and is described in Fig. 2.

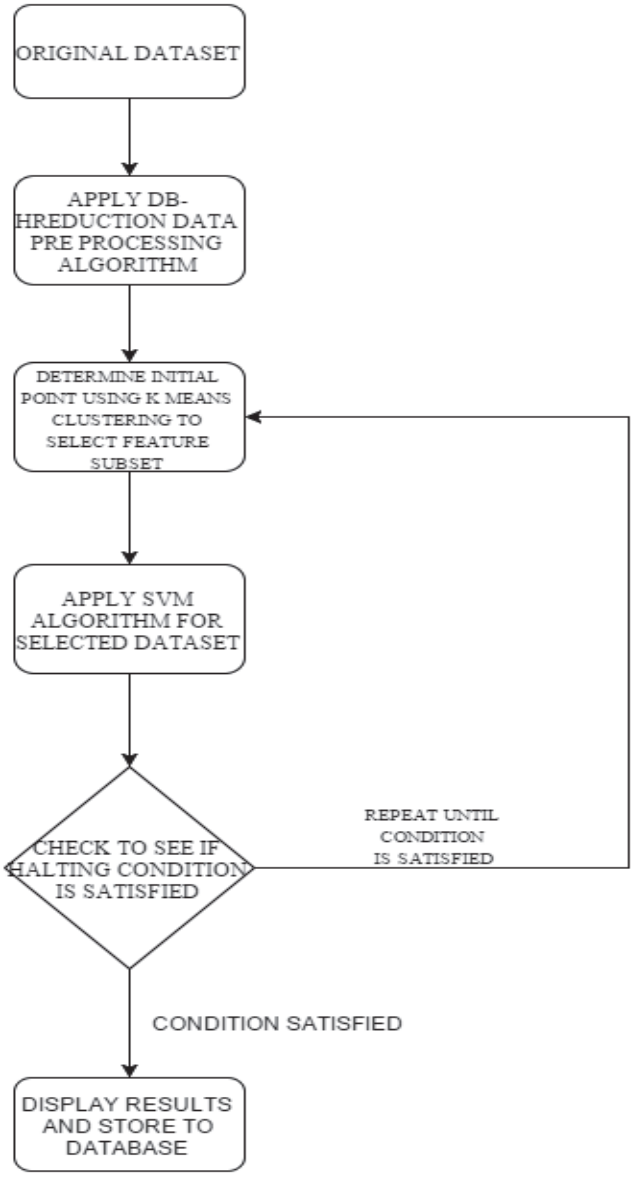

Figure 2. Flowchart of the proposed heuristic model

A. Original Dataset

The original dataset consists of the raw pool of input data which may contain noise, may be inconsistent and incomplete. The original dataset is therefore not suitable for applying any clustering algorithm as it may lead to inconsistent and random clusters. The raw dataset therefore needs to be preprocessed before applying any clustering or SVM algorithm. The original dataset contains the data in Fixed Column Format (FCF)

B. Data Preprocessing

In the second stage of the model, the data is preprocessed before making it ready for applying algorithms and performing operations. The sub-tasks performed in this stage are a) Data Cleaning b) Data Transformation c) Data Integration d) Data Discretization and e) Data Reduction [18].

In Data Cleaning sub-stage, the tuples without the class label is ignored and replaced with the mean (μ) of the attributes. Next, the noisy data is cleaned using the process of binning where the data is portioned into divisions called bins and then smoothed using the median of those bins. Inconsistent data is removed as per the data's type. In the Data Transformation sub-phase, the data is normalized to fall within the range $\mathcal{R} \in [min, max]$. Data Aggregation may not be required if numerical attributes are absent. In the Data Integration sub-stage, duplicate data is merged to form a single datum. Data and metadata conflicts are resolved by correlational analysis. The Data Discretization sub-stage converts the data into discrete data by dividing the range of a continuous attribute into intervals and reduces the data size. The Data Reduction sub-stage removes the irrelevant attributes and reduces the number of tuples by sampling. Fig. 3 shows how data travels from the input database to the output database.

Now, the processed data is ready to be used for applying a cluster algorithm.

C. Selecting Feature Subset using K-Means Clustering

K-Means clustering algorithm as described in Tan et al. [19] is used select the feature subset in the dataset. In this algorithm, the clustering approach divides the dataset into partitions. Each cluster has a center point which is assigned to nearest cluster. The distance is measured by the Euclidean distance [20] and is given by:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Initially, K points are chosen and after assigning all the points to the nearest cluster, the centroid of each cluster is continuously calculated until the centroids remain constant. After K-Means clustering is applied, the initial starting point is chosen for evaluating the feature selection subset.

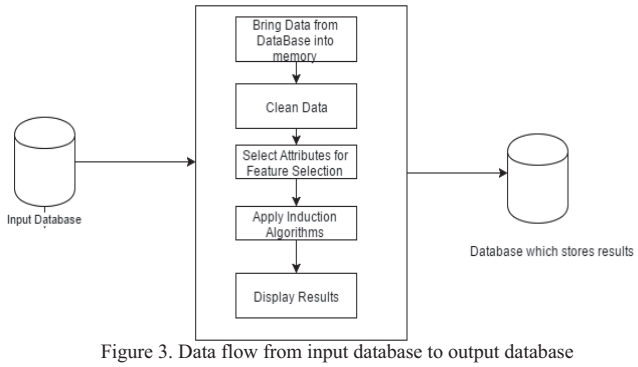

Figure 3. Data flow from input database to output database

D. Applying SVM Algorithm

The clustered dataset is then subjected to Support Vector Machine (SVM) classifier which optimally maps $\mathcal{X} \mapsto \mathcal{Y}$, where $x \in \mathcal{X}$ is an object and $y \in \mathcal{Y}$ is a class label. The algorithm yields a hyper-plane which maps one cluster with another. This gives the largest minimum distance to the training dataset. It is important for the dataset to be noise free to get optimal results.

The support vector obtained is the hyperplane which is closest to the training example. If the dataset is non-linear, the misclassification errors need to be minimized by modifying the optimization model.

After obtaining the support vectors, the dataset is correctly classified and grouped according to the parameters and data types.

E. Checking for Halting Condition

In this phase, an iterative approach is followed where the halting condition is checked. If the end of the complete search space is reached or the clustering accuracy increases to a significant level, the algorithm is stopped. If the accuracy decreases or some attributes are still left to be clustered, the K-Means clustering is applied again followed by SVM algorithm.

F. Generating the Resultant Dataset

The resultant dataset contains an optimally clustered subset with definitive characteristic features. The resultant dataset can be used to generate patterns of data. The clustered data contains their own characteristics and the information can be easily extracted using simple techniques. The resultant dataset is clean and any further data can be classified easily into existing clusters or new clusters.

V. Conclusion

Machine Learning Learning techniques can effectively aid Data Mining. Feature Selection, Fuzzy Method and Decision Trees can be used to accurately cluster and model a large pool of data and also make room for predicting the classifier functions for future datasets. The paper introduced a heuristic model and followed feature selection to cluster a large dataset. It was found that Forward and Backward Multiple Regression is an ineffective approach to determine the starting point for feature selection. Moreover, SFS and SBS approaches are the best with Forward Selection being the preferred method. The heuristic model proposed uses a SVM because of it's accuracy despite being a bit computationally heavy. The dataset will be able to determine the level of accuracy of SVM.

Using Machine Learning in Data Mining can aid in many real life problems such as facial recognition in social networking data. It is extensively being used in Biological data analysis to cluster species and can improve industries to manage Big Data well. Data Mining is further used in retail, stock market analysis, banking, weather forecast, and medical industries. The main applications of feature selection are drug screening, identifying cancer cells, text filtering etc. Future work includes implementing the heuristic model with real life datasets with differential characteristics to study its performance more effectively.

References

[1] Mitchell, Tom; Carbonell, James; Michalski, Ryszard (1986). "Machine Learning: A Guide to Current Research (The Springer International Series in Engineering and Computer Science)".

[2] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) Foundations of Machine Learning, The MIT Press.

[3] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases".

[4] Kira, K., Rendell, L.A., 1992. A practical approach to feature selection. In: Proceedings of the Ninth International Conference on Machine Learning. pp. 249–256.

[5] Meisel, W.S., 1972. Computer-Oriented Approaches to Pattern Recognition. Academic Press, New York.

[6] Selwyn Piramuthu, Evaluating feature selection methods for learning in data mining applications, European Journal of Operational Research, Volume 156, Issue 2, 16 July 2004, Pages 483-494, ISSN 0377-2217.

[7] Eyke Hüllermeier, Fuzzy methods in machine learning and data mining: Status and prospects, Fuzzy Sets and Systems, Volume 156, Issue 3, 16 December 2005, Pages 387-406, ISSN 0165-0114.

[8] Quinlan, J.R., 1987. Simplifying decision trees. International Journal of Man–Machine Studies 27, 221–234.

[9] Josef Kittler, A nonlinear distance metric criterion for feature selection in the measurement space, Information Sciences, Volume 9, Issue 4, 1975, Pages 359-363, ISSN 0020-0255.

[10] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

[11] Chidanand Apté, Sholom Weiss, Data mining with decision trees and decision rules, Future Generation Computer Systems, Volume 13, Issues 2–3, November 1997, Pages 197-210, ISSN 0167-739X.

[12] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Monterrey, CA., 1984.

[13] Avrim L. Blum, Pat Langley, Selection of relevant features and examples in machine learning, Artificial Intelligence, Volume 97, Issues 1–2, December 1997, Pages 245-271, ISSN 0004-3702.

[14] I.M. Mujtaba, M.F. Hussain, Application of Neural Networks and Other Learning Technologies in Process Engineering, 2001, Imperial College Press, ch. 2, pp. 23-48.

[15] D.D. Jensen, P.R. Cohen, Multiple Comparisons in Induction Algorithms, Machine Learning, March 2000, Volume 38, Issue 3, pp 309-338.

[16] A. Abraham, L. Jain, B.J. van der Zwaag, Innovations in Intelligent Systems, 2004, Studies in Fuzziness and Soft Computing (Springer), ch. 1, pp. 3-20.

[17] L. Breiman, Random Forests, Machine Learning, October 2001, Volume 45, Issue 1, pp. 5-32.

[18] S. García, J. Luengo, F. Herrera, Data Preprocessing in Data Mining, 2015, Intelligence Systems Reference Library (Springer), ch. 3-6, pp. 39-162.

[19] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[20] M.M. Deza, E. Deza, Encyclopedia of Distances, 2009, Springer, pp. 94.