ISSN: 0974-5823 Vol. 6 No. 3 December, 2021

International Journal of Mechanical Engineering

EFFECTIVENESS OF INFORMATION THEORY IN MACHINE LEARNING-BASED PREDICTION AND CLASSIFICATION FOR CLINICAL AND MECHANICAL PROBLEMS – A COMBINATIONAL SURVEY

[1]Murugan R, [2]Senbagamalar L,

[1]School of Computer Science and IT

[2]Computer Science Engineering

[1]Jain Deemed to be University, Jayanagar, Bangalore, Karnataka, India.

[2]Bannari Amman Institute of Technology, Sathyamangalam, TamilNadu, India.

Abstract

Machine learning algorithms that are all automated and effective in identifying the underlying pattern from the available data have gained their importance in almost all types of real-time applications. The application area of machine learning methods includes the clinical, production, stock market, defense, agriculture, education, etc. Two significant operations performed by machine learning algorithms are prediction and classification. Prediction is a methodology to identify how could the data pattern travel in the future and the classes can be accommodated. In contrast, classification is to categorize the classes for the available data. There are specific application areas where prediction, as well as classification, are carried out simultaneously. The machine learning algorithms give better results only when the data is interpreted better, which can be done through feature engineering, where information theory plays a vital role. Statistical information theory is implemented in many machine learning applications using three approaches, namely entropy, information gain, and mutual information. In this article, we carried out the combinational survey of the machine learning approaches, which implemented the information theory for feature analysis in two major areas, namely clinical and mechanical. The survey could guide future researchers to get a clearer idea of the statistical approaches to adopt for the valuable outcome.

Keywords: Prediction, Classification, Information Gain (IG), Mutual Information (MI), Statistics and Entropy.

1. Introduction

In the everyday activities of a person, the enormous amount of irrelevant data or information might lead to the failure of the person's significant tasks. Let us consider a real-time scenario where the usage of a car without the rear mirror and the use of mobile phones while driving such a car. In this scenario, the person driving the vehicle is vulnerable to an accident. Artificial Intelligence also faces a similar problem while dealing with irrelevant information. Machine learning algorithms that effectively classify and predict the underlying pattern from the available data might mislead if an enormous amount of irrelevant information lies within the data. Thus it becomes essential to identify the abstract data set to obtain high-performance machine learning models. Feature engineering is an important area in machine learning where the basic set of features is computed, and the best location is obtained using various statistical measures. The data considered for analysis is preprocessed and feature calculated data. Then it is enough to select or extract the compact set using the available selection and extraction machine learning methods. Suppose we have a raw collection of data, i.e., image data. In that case, it has to be preprocessed using image processing techniques, and essential features to describe the data have to be computed. The feature selection or extraction strategies for such raw data are only possible after identifying the relevant parts. A similar implementation strategy can be adopted for data, including signal, video, and audio types. All such raw data has to be preprocessed, and the essential feature must be computed for better interpretation through the data. As mentioned earlier, statistical measures must be adopted for better performance during the computation, selection, and extraction stages. The transformation of the dataset from its available form to a better interpretable form is also done through statistical methods. Several statistical methods are utilized for data interpretation among those based on the efficiency and implementation stages considered for analysis; three different statistical methods are considered for the survey, namely Mutual Information (MI), Information Gain (IG), and Entropy. A machine learning model needs to quantify the number of information received for attaining or performing any tasks. Information theory is a mathematical concept that can be adopted for computational implementation, and better results can be obtained for the available data. In the feature, computational stage entropy and its different mathematical variants available will support the construct of the dominant elements in describing the data and its dimension. Information Gain and Mutual Information are beneficial in any two stages considered for combining the feature set, i.e., feature selection

or feature extraction. Feature selection from a machine learning perspective is the area that deals with selecting the appropriate attributes that play a vital role in describing the data. Feature extraction transforms the data from one dimension to another so that the interpretability of the data might get improved. Selection of attributes can be made through the filter, wrapper, and embedded approaches. It is possible to adopt the statistical methods in any of these three different methods, or it can be assumed in an ensemble mode so that the feature selection will produce a compact set of features for analysis. Feature extraction can be done through methods like Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), and so on. While reducing the dimension of the data, it is essential to consider the statistical impact of the features so that the components can be ordered.

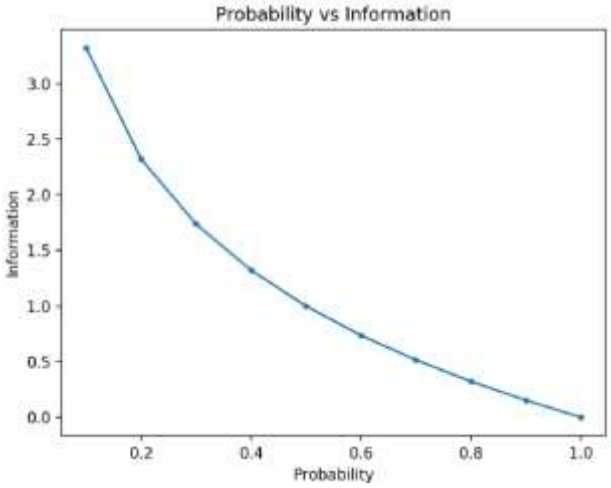

Figure 1: Plot for Probability of Random Variables vs. Information

While considering the data for statistical analysis, it is also crucial to analyze the data type since the distribution and distance calculation is the primary operations performed by the statistical methods. If the data has static distribution, it becomes easier for the available algorithms to select the relevant features. In contrast, if the data has a lot of random variables, it will be difficult for an algorithm to find the feature subset which requires any of the statistical methods for assistance. Figure 1 shows how the plot of probabilistic random variables differs in every iteration regarding the data considered for analysis in each iteration provided with random variables.

In figure 1, the probability range has been set from 0 to 1, and for information, Euler's numbers up to 3 are considered for processing. From the figure, it is evitable that probability fluctuates based on the information given to the model. Thus, we framed the article's flow by considering the requirement for information theory in machine learning implementation models. Section 2 denotes the motive of the survey, which describes the number of researches and the methods implemented in machine learning applications along with the statistical methods. Section 3 represents the outcome obtained through the survey through the comparison and the results from the existing works, and section 4 concludes the study.

2. Survey Motivation

Entropy is a mathematical theory that handles the uncertainty problem of the data so that the repeated and irrelevant information can be eradicated from the processing stage, which maintains the certainty for the available data [1]. While entirely concentrating on the specific nature of the data, there is a possibility of increased complexity in time, which is a primary concern while dealing with an enormous amount of data [2]. Entropy-based feature selection was adopted to handle the noise and classify the clinical problem, i.e., differentiate between normal and abnormal Vibroarthrographic (VAG) signal samples [3]. Since the classification of clinical issues is gaining importance, we considered the number of clinical problem analyses in our survey. In the feature computing stage from the VAG signals, categories including Entropy, Shannon entropy (SHE), and Multiscale Sample Entropy (MSE) are computed for identifying the best of features that describe the underlying pattern of the data [4]. Entropy is also utilized as a feature computation and multiclass classification parameter so that the different levels involved in clinical pathologies can be sorted out [5]. Entropy can be computed for each sample considered for analysis and based on the entropy value to quantify the data. The ranking scores the best set of features are identified and denoted as feature subset for classification [6]. The entropy calculation after the feature selection is considered to improve the overall prediction range of the data and to validate the individual certainty of the selected set of samples [7]. The essentiality of feature study and the clearer idea regarding the feature strategy to be adopted, i.e., carrying with selection or extraction, is the critical focus area while dealing with machine learning-based classification and prediction problems [8]. While dealing with multiple classification problems, computation of entropy-based features will improve the time–domain information so that the classification range and the data point at which the class is to be categorized can be identified better [9].

Mutual information is another critical statistical parameter that could identify the interrelationship within the available features. Searching efficiency and relevance matching are the two essential tasks while performing the feature selection where relevance plays a vital role in identifying the best set of features since it calculates the dependability of one attribute or part with the other one [10]. Thoracic data analysis using Artificial Neural Network (ANN) had been carried out. The initial step from the available features in the database is to calculate the features' Mutual Information (MI). Once the MI is computed, it is set to the limits using the joint and marginal probability distributions, which had further scrutinized the feature subset to be more relevant and dependent [11]. The same cancer dataset is utilized. The subset of the features is further interpreted by calculating relevance-assisted feature selection and classification methodology, which also improved clinical data analysis performance [12]. The two major stages available in the feature selection approaches are filtered and embedded methods since both can perform the combinatory analysis. In the filter approach, it is possible to implement the mutual information strategy since it identifies the most relevant set of features. Further filtering using an embedded system will improve the machine learning model [13]. Ensemble methods combine the better criterion available in each way adopted for clinical analysis by calculating the relevance between the attributes and calculating the subset from the original dataset considered for analysis [14]. The claim that relevant information will be more beneficial for classification problems must be justified only based on the number of machine-based implementation algorithms that the model had gone through and the best results produced. The ensemble of different ML

algorithms is done through the simulations, and the results are based on a comparison strategy is adopted [15].

The essential parameter to notify the impact of the data while performing machine-based classification or prediction is Information Gain (IG). Information Gain supports retrieving the valuable and additive information from the design space of the mechanical devices to be developed to provide high performance compared with the available ones [16]. ML algorithms, including Linear Regression, Logistic Regression, Support Vector Machine (SVM), and Naïve Bayes classifier are utilized to test the wear loss, and the complete information regarding the wear loss occurred in the alloy is evaluated using the Information Gain (IG) which had improved the overall prediction rate [17]. Surface roughness level calculation from the mechanical data describes the strength and durability of an alloy which can be done by calculating the maximum amount of surface relevant information that is directly calculated from the dataset, and the feature engineering strategy along with classification can be adopted for categorizing the surface-oriented alloy based information [18]. Identification of roughness in a specific surface can be made by combining Artificial Neural Networks (ANN) and Logistic Regression. The prediction of roughness and smoothness data using the available features regarding the data is the essential parameter. This parameter is satisfied by calculating the features' information gain during the selection phase [19]. The error rate of the fatigue life of different categories of alloys can be predicted using machine learning strategies where knowledge-based systems play a vital role. In such a scenario, complete information obtained regarding the dimensions and features of the alloy acts as a knowledge-based system that provides valuable resources to the ML classifiers [20].

3. Survey Outcome

The different eminent research areas are considered for the statistical implementation methods involved in machine learning techniques. Three major statistical information concepts, namely Entropy, Mutual Information, and Information Gain, are considered implementation measures. The considered measures have been implemented in many areas. We had identified two major areas: clinical assessments (i.e., for Entropy and Mutual Information) and mechanical production in alloys (for Information Gain). As mentioned earlier, we considered almost 18 major works done using the statistical concepts in two significant areas.

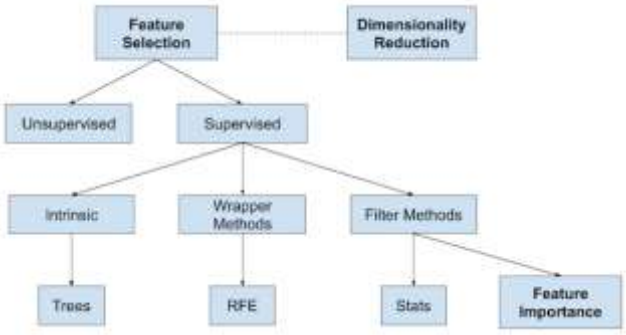

Figure 2: Overview of Feature Engineering stages

The survey outcome had also split into three primary information methods so that it will be helpful for better understandability of the impact of statistics in feature engineering and machine learning. The implementation stages involved in the statistical implementation of machine learning methods are shown in figure 2. The initial outcome obtained by six major research works and its outcomes were considered to project the impact of using entropy in clinical predictions. The analyzed works have utilized a common benchmark dataset with 89 VAG signal samples, 41 attributes representing the various dimensions of the dataset are being used. The outcome is compared through the accuracy, sensitivity, and specificity results obtained. Table 1 compares the entropy utilized in machine learning models and their effects.

Table 1: Performance Comparison of Entropy-Based Feature Selection

Author	Method	Accuracy	Sensitivity	Specificity
Nalband et al. (2018)	Ensemble-based Feature selection and Decomposition	87	91	79
KręciszK (2018)	Multiclass Classification based on Entropy features	90	93	84
Wu.et.al (2018)	Entropy feature calculation and envelope for amplitude measures.	82	85	72
Befrui et al. (2018)	Normalized entropy feature computation and classification using SVM	89	80	75
Alphonse Balajee et. al (2021)	Modeling and Multiclass classification using divergence Random Forest	82	75	53
BalajeeA (2021)	Greedy Regressive informative Feature selection	97	98	83

There are two different types of classification strategies adopted in the literature considered for analysis: binary classification and multiclass classification. The results are compared commonly based on the entropy feature selection, not in the type of classification method adopted. The graphical

representation of the comparison of the results is shown in figure 3.

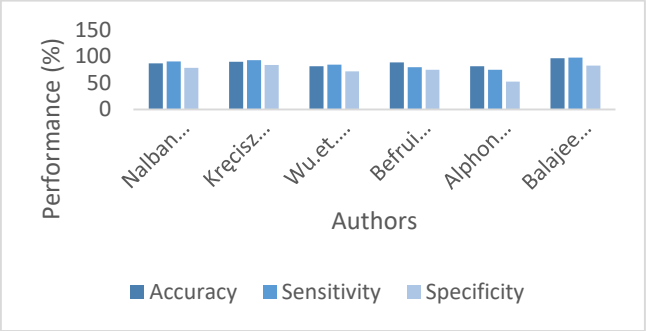

Figure 3: Graphical Representation of Entropy-Based Performance Comparison

The analysis of mutual information as a statistical parameter to perform better feature selection is considered the second outcome comparison strategy. The second clinical problem thought an application area for the literature is the classification of samples to check whether an individual is affected by heart disease. The thoracic dataset is commonly used by all the literature considered for the survey. The performance comparison of the results obtained for the methods that had used mutual information-based feature selection is shown in table 2.

Table 2: Performance Comparison of Mutual Information-Based Feature Selection

Author	Method	Accuracy	Sensitivity	Specificity
Podolsky et al. (2016)	Gene Expression Level Prediction using Logistic Regression	84	78	72
Das et al. (2017)	Bi-Objective feature selection genetic algorithm	82	87	79
Azumi et.al(2019)	Boosted Neural Network along with Ensemble Classifier	93	91	92
Ganglia et.al(2021)	Hybridization of mutual informative filter and wrapper approaches	92	79	84
Thaventhiran C and Sekar K R (2021)	Feature Matching and DNN for Prediction	98	96	93

Clinical samples and their interpretation using machine learning algorithms will identify the underlying pattern of the data, which could be more useful in prediction and classification problems. In our survey. We had considered heart disease prediction and classification of samples into normal and abnormal classes. While calculating the feature impact using the matching scores in each iteration, it will be helpful for the model to perform better prediction using the machine learning and deep learning approaches [13]. The graphical representation of the results obtained for mutual information calculated feature selection are given in figure 4

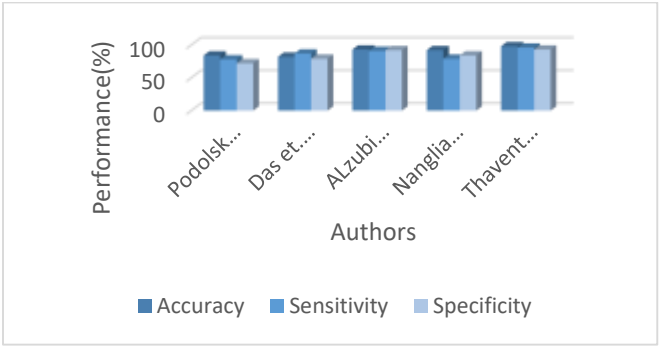

Figure 4: Graphical Representation of Mutual Information Based Performance Comparison

In the production industry, wholly focused on machinery implementations, alloy production has gained paramount importance due to its effective and comprehensive utilization. The major challenge faced by the production industry is to calculate the roughness of the alloy so that the field towards which the alloy to be used might get changed.

Table 3: Performance Comparison of Information Gain-Based Feature Selection

Author	Method	Accuracy	Sensitivity	Specificity
Çaydaş U, Hascalık A (2008)	Roughness calculation using ANN	80	77	63
Altay et al. (2020)	Wear loss prediction in ferroalloy coating using Singular Value Decomposition	84	91	86
Ulas et al. (2020)	Surface Roughness level prediction using LS-SVM	91	83	89
Lian et al. (2022)	Lifetime prediction of alloy using LSTM	97	92	98

Manual calculation of roughness might be erroneous and time-consuming, where the need for machine learning algorithms becomes an essential tool for application. We considered four primary research methods that all had utilized different information gain-based feature selection strategies for calculating the roughness of an alloy. The comparison of results for the considered research methods is shown in table 3. The recent research work had gained better performance, which utilized LSTM for calculating the loss and gain of the roughness in the alloy surface. The graphical representation of the results obtained is shown in figure 5.

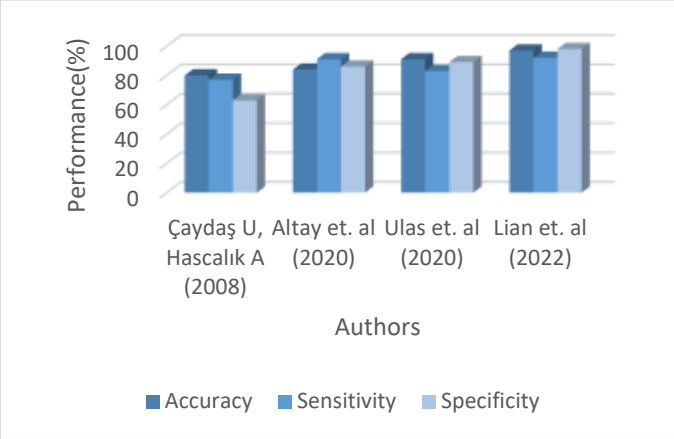

Figure 5: Graphical Representation of Information Gain Based Performance Comparison

4. Conclusion

The survey's primary focus is to analyze the impact of statistical methods used as an essential parameter to perform feature computation, extraction, and selection methods. The different applications had utilized feature engineering and classification methods which had given a better performance. Two major application areas, namely medical diagnosis and mechanical production, are considered, and the analysis of machine learning methods and the outcome were analyzed in this survey. This could provide a combinatory analysis to the data scientist who is working in the machine learning-based application areas to develop a better model that could improve the performance. At the same time, compared to all the other existing methods.

References

[1] C.E. Shannon, A mathematical theory of communication, SIGMOBILE Mobile Comput. Commun. Rev. 5 (1) (2001) 3–55.

[2] Y.H.Qian, J. Y.Liang, F.Wang, A new method for measuring the uncertainty in incomplete information systems, Int.J. Uncertain. Fuzziness Knowl-Based Syst. 17(6) (2009)855–880.

[3] Nalband S, Prince A, Agrawal A. Entropy-based feature extraction and classification of vibroarthographic signal using complete ensemble empirical mode decomposition with adaptive noise. IET Science, Measurement & Technology. 2018 Apr 26;12(3):350-9.

[4] Balajee A, Venkatesan R. Machine learning based identification and classification of disorders in human knee joint–computational approach. Soft Computing. 2021 Oct;25(20):13001-13.

[5] Kręcisz K, Bączkowicz D. Analysis and multiclass classification of pathological knee joints using vibroarthrographic signals. Computer methods and programs in biomedicine. 2018 Feb 1;154:37-44.

[6] Wu Y, Chen P, Luo X, Huang H, Liao L, Yao Y, Wu M, Rangayyan RM. Quantification of knee vibroarthrographic signal irregularity associated with patellofemoral joint cartilage pathology based on entropy and envelope amplitude measures. Computer methods and programs in biomedicine. 2016 Jul 1;130:1-2.

[7] Befrui N, Elsner J, Flesser A, Huvanandana J, Jarrousse O, Le TN, Müller M, Schulze WH, Taing S, Weidert S. Vibroarthrography for early detection of knee osteoarthritis using normalized frequency features. Medical & biological engineering & computing. 2018 Aug;56(8):1499-514.

[8] Balajee A, Venkatesan R. A survey on classification methodologies utilized for classifying the knee joint disorder levels using vibroarthrographic signals. Materials Today: Proceedings. 2021 Jul 26.

[9] Alphonse Balajee, et al. "Modeling and multi-class classification of vibroarthographic signals via time domain curvilinear divergence random forest." Journal of Ambient Intelligence and Humanized Computing (2021): 1-13.

[10] Gao L, Wu W. Relevance assignation feature selection method based on mutual information for machine learning. Knowledge-Based Systems. 2020 Dec 17;209:106439.

[11] ALzubi JA, Bharathikannan B, Tanwar S, Manikandan R, Khanna A, Thaventhiran C. Boosted neural network ensemble classification for lung cancer disease diagnosis. Applied Soft Computing. 2019 Jul 1;80:579-91.

[12] Nanglia P, Kumar S, Mahajan AN, Singh P, Rathee D. A hybrid algorithm for lung cancer classification using SVM and Neural Networks. ICT Express. 2021 Sep 1;7(3):335-41.

[13] Thaventhiran C, Sekar KR. Target Projection Feature Matching Based Deep ANN with LSTM for Lung Cancer Prediction. INTELLIGENT AUTOMATION AND SOFT COMPUTING. 2022 Jan 1;31(1):495-506.

[14] Das AK, Das S, Ghosh A. Ensemble feature selection using bi-objective genetic algorithm. Knowledge-Based Systems. 2017 May 1;123:116-27.

[15] Podolsky MD, Barchuk AA, Kuznetcov VI, Gusarova NF, Gaidukov VS, Tarakanov SA. Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. Asian Pacific journal of cancer prevention. 2016;17(2):835-8.

[16] Geetha NK, Bridjesh P. Overview of machine learning and its adaptability in mechanical engineering. Materials Today: Proceedings. 2020 Nov 5.

[17] Altay O, Gurgenc T, Ulas M, Özel C. Prediction of wear loss quantities of ferroalloy coating using different machine learning algorithms. Friction. 2020 Feb;8(1):107-14.

[18] Ulas M, Aydur O, Gurgenc T, Ozel C. Surface roughness prediction of machined aluminum alloy with wire electrical discharge machining by different

machine learning algorithms. Journal of Materials Research and Technology. 2020 Nov 1;9(6):12512-24.
[19] Çaydaş U, Hascalık A. A study on surface roughness in abrasive waterjet machining process using artificial neural networks and regression analysis method. Journal of materials processing technology. 2008 Jun 20;202(1-3):574-82.
[20] Lian Z, Li M, Lu W. Fatigue life prediction of aluminum alloy via knowledge-based machine learning. International Journal of Fatigue. 2022 Jan 3:106716.