International Journal of Computational Intelligence in Control

EVALUATION OF FEATURE SELECTION METHODS FOR DISEASE SURVIVAL PREDICTION USING PROPOSED ENTROPY-BASED FUZZY TOPSIS TECHNIQUE

Dr. M. LALLI[1], S. AMUTHA[2]

[1]Assistant Professor in Computer Science, Bharathidasan University, Tiruchirappalli, Tamilnadu.

[2]Research Scholar in Computer Science, Bharathidasan University, Tiruchirappalli, Tamilnadu.

Date of Submission:15th March 2021 Revised: 29th May 2021 Accepted: 17th June 2021

Abstract; The medical industry is making strides through modern technical advances and the development of newer healthcare and treatment techniques. Biotechnology is the basis of all these technical advancements. The health of individuals is declining each day with the introduction of many toxins, materials and chemicals in our everyday lives. It affects not only physical or psychological health, but also our lifestyle. The type and stage of skin cancer treatment depends on the size and location of the tumour, and the overall health and medical history of your patient. The purpose of treatment in most cases is to complete cancer removal or destruction. If detected and treated early, the majority of diseases can be cured.This paper is aim to evaluate the feature selection technique based on the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). An enhanced Entropy-based TOPSIS method is developed to suggest the one or more choices among alternatives, having many attributes.The proposed TOPSIS method is used to analyze the performance of the feature selection to enhance the skin disease classification.

KEYWORDS: Skin cancer, Machine Learning, Multi Criteria Decision Making, Technique for Order of Preference by Similarity to Ideal Solution, Feature Selection

1. INTRODUCTION

Healthcare has long been a data-rich field. With so many moving parts, healthcare providers and insurers have no shortage of variables to measure. The captured data have many important uses [4]. They keep tabs on costs and billing. They track the activity of hospitals and outpatient facilities. Crucially, the data record the health states of people at a microscopic and macroscopic level. It is hard to overstate the importance of data in healthcare, especially when it comes to improving healthcare systems.

Almost 61% of deaths in India are presently ascribed to Non-communicable Diseases (NCD), including heart issue, cancer and diabetes, as indicated by new information discharged by the World Health Organization on Monday. Very nearly 23% are in danger of premature death because of such ailments. In India, an aggregate of 58,17,000 deaths were evaluated from illnesses like cancer, diabetes and heart issues in 2016. Cardiovascular infections (coronary illness, stroke, and hypertension) add to 45% of all NCD deaths, trailed by chronic respiratory disease (22%), cancer (12%) and diabetes (3%). Cancer, Diabetes and heart disease alone record for 55% of the untimely mortality in India in the age gathering of 30-69 years. With the change of expectations for everyday comforts, the occurrence of chronic disease is expanding. It is basic to perform hazard appraisals for chronic disease. With the development in therapeutic information, gathering Electronic Health Records (EHR) is progressively advantageous. Patients' factual data, test results and infection history are recorded in the EHR, empowering us to distinguish potential answers for lessen the expenses of restorative contextual analyses [5].

2. RELATED WORKS

Liang, Huiying, et al [3] showed that Machine Learning Classifiers (MLCs) can query EHRs in a manner similar to the hypothetico-deductive reasoning used by physicians and unearth associations that previous statistical methods have not found. Our model applies an automated natural language processing system using deep learning techniques to extract clinically relevant information from EHRs.

Wu, Chieh-Chen, et al [4] aimed to develop a machine learning model to predict Fatty

Liver Disease (FLD) that could assist physicians in classifying high-risk patients and make a novel diagnosis, prevent and manage FLD. Classification models such as random forest (RF), Naïve Bayes (NB), artificial neural networks (ANN), and logistic regression (LR) were developed to predict FLD.

Jo, Taeho, Kwangsik Nho, and Andrew J. Saykin [5] A systematic review of publications using deep learning approaches and neuro imaging data for diagnostic classification of Alzheimer's disease (AD) was performed. A PubMed and Google Scholar search was used to identify deep learning papers on AD published between January2013 and July 2018. These papers were reviewed, evaluated, and classified by algorithm and neuro imaging type, and the findings were summarized.

Ngiam, Kee Yuan, and Wei Khor [6] analyzed big data by machine learning offers considerable advantages for assimilation and evaluation of largeamounts of complex health-care data. However, to effectively use machine learning tools in health care, several limitations must be addressed and key issues considered, such as its clinical implementation and ethics inhealthcaredelivery. In this Review, the authors discuss some of the benefits and challenges of big data and machine learning in health care.

Kawakami, Eiryo, et al [7] aimed to develop an ovarian cancer–specific predictive framework for clinical stage, histotype, residualtumor burden, and prognosis using machine learning methodsbased on multiple biomarkers. Machine learning systems can provide critical diagnostic and prognostic prediction for patients with EOC before initial intervention, and the use of predictive algorithms may facilitate personalized treatment options through pre-treatment stratification of patients.

Abdar, Moloud, et al [8] describe an innovative machine learning methodology that enables an accurate detection of Coronary Artery Disease (CAD) and apply it to data collected from Iranian patients. The authors first tested ten traditional machine learning algorithms, and then the three-best performing algorithms (three types of SVM) were used in the rest of the study. To improve the performance of these algorithms, a data pre-processing with normalization was carried out. Moreover, a genetic algorithm and particle swarm optimization, coupled with stratified 10-fold cross-validation, were used twice: for optimization of classifier parameters and for parallel selection of features.

3. IMPORTANCE OF FEATURE SELECTION IN PRE-PROCESSING

Several, feature selection techniques have proposed in the paper, and their comparative examination is a challenging task [9]. Without identifying the appropriate attributes in an approach of the live dataset, it is complicated to determine the strength of the feature selection techniques, because datasets may incorporate several hurdles such as the vast number of redundant and irrelevant, high dimensionality and noisy data in term of samples or features. Hence, the efficiency of the feature selection technique swears on the learning approach performance. There are several performance criteria discussed in the literature such as the ratio of feature selection, accuracy, error rates, etc. Most researchers accept that there is no so-called "trustworthy system." Therefore, the new feature selection techniques are continually developing to undertake the particular difficulty with various procedures. Utilizing an ensemble method for assuring a reliable performance of feature selection [9,10].

4. PROPOSED FRAMEWORK FOR EVALUATING FEATURE SELECTION TECHNIQUES WITH PROPOSED ENTROPY BASED TOPSIS METHOD

The following figure 1 represents the framework for evaluating the feature selection technique by using Classification methods for given disease survival dataset [12]. The proposed Entropy-based TOPSIS method is used to estimate the feature selection technique.

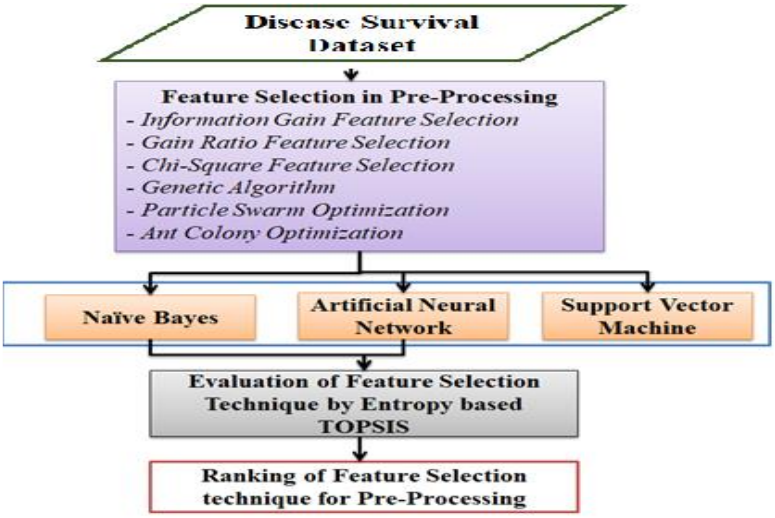
Disease Survival Dataset
Feature Selection in Pre-Processing
- Information Gain Feature Selection
- Gain Ratio Feature Selection
- Chi-Square Feature Selection
- Genetic Algorithm
- Particle Swarm Optimization
- Ant Colony Optimization
Naïve Bayes
Artificial Neural Network
Support Vector Machine
Evaluation of Feature Selection Technique by Entropy based TOPSIS
Ranking of Feature Selection technique for Pre-Processing

Figure 1: Framework for the evaluation of Feature Selection techniques by using proposed Entropy based TOPSIS

4.1 Information Gain Feature Selection

Given the entropy is a condition of impurity in a training set S, it can describe a measure excogitating additional information about Y rendered by X that represents the amount by which the entropy of Y decreases. This measure is known as IG. It is given by

$$IG=H(Y)-H(Y|X)=H(X)-H(X|Y) \quad (3)$$

IG is an asymmetrical pattern (refer to equation 3)). The information gained about Y after examining X is equal to the information gained about X after scrutinizing Y. A delicacy of the IG measure is that it is predetermined in support of features with high values even when they are not more informative.

4.2 Gain Ratio Feature Selection Technique

The Gain Ratio [8] is the non-symmetrical measure that is introduced to compensate for the bias of the Information Gain (IG) [7]. GR is given by

$$GR = \frac{Information\ Gain\ (IG)}{H(X)} \quad (3)$$

Information Gain (IG) is an equal measure.

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y) \quad (4)$$

The information gained about Y after scrutinizing X is similar to the information gained about X after examining Y. A delicacy of the IG measure is that it is predetermined in support of features with higher values even when they are not more informative.

As in the equation (3) presents, when the variable Y has prognosticated, then normalize the IG by splitting by the entropy of *X*, and vice versa. Because of this normalization, the GR values constantly come in the range [0, 1]. A value of GR = 1 indicates that the knowledge of *X* completely predicts *Y*, and GR = 0 means that there is no relation between *Y* and *X*. In opposition to IG, the GR favors variables with fewer values.

4.3 Chi-Square Feature Selection Method

Feature Selection via chi-square χ^2 test is another, very commonly used the method. Chi-squared attribute evaluation evaluates the worth of a feature by computing the value of the chi-squared statistic concerning the class. The initial hypothesis H_0 is the assumption that the two features are unrelated, and it is tested by the chi-squared formula:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{O_{ij}-E_{ij}}{E_{ij}}\right)^2 \quad (4)$$

Where O_{ij} is the observed frequency, and E_{ij} is the expected (theoretical) frequency, asserted by the null hypothesis. The greater the value of χ^2, the greater the evidence against the hypothesis H_0

4.4 Genetic Algorithm based Feature Selection Method

Genetic algorithms (GA) are an adaptive heuristic search method based on the idea of natural selection [11]. They are inspired by Darwin's theory of evolution – "survival of the fittest", which is one of the randomized search techniques.The algorithm begins with a set of individuals (chromosomes) called as population. Individual chromosome consists of a set of genes that could be bits, numbers or characters. Individuals are selected according to their fitness value for reproduction. Higher the fitness value more is the chances of an individual being selected. Crossover and mutation is responsible for producing new population. Crossover accelerates the search early in the evolution of the population, while mutation is responsible for restoring the lost information to population by local or global movement in the search space. The process is iteratively repeated several times until stopping criteria are met or optimal solution is reached.

4.5 Particle Swarm Optimization

PSO has adjusted with a cluster of random particles (solutions). The procedure then hunts for optima by a sequence of iterations. The particle's relevance value has estimated on every iteration. If it is the known value the particle has obtained, the particle stores the position of that condition as pbest (particle best). The position of the best fitness value obtained by any particle when any iteration has saved as gbest (global best). By using gbest and pbest, each particle goes with a specific velocity, calculated by the following equation:

$$V_i = wV_{i-1} + c_1 * rand() * (pbest - pL) + c_2 * rand() * (\text{gbest-pL}) \quad (1)$$

$$pL = pvL + V_i \quad (2)$$

$$w = \frac{1}{iterNum} \quad (3)$$

The present velocity has presented by Vi. V_{i-1} is the past velocity, and pL is the present position of the particle. pvL is the previous location of the particle and rnd is a random number between (0, 1). The learning factors are c1 and c2 or stochastic factors, and the present iteration number is iter Num.

4.6 Ant Colony Optimization Feature Selection Technique

Ant colony optimization has introduced by Marco Dorigo and his colleagues in the early 1990s. The first computational model appeared by the name ant system (AS). It is another strategy to stochastic combinatorial optimization. The search activities have distributed over: "ants" – agents with natural basic inclinations that simulate the behavior of real ants. The main objective was not to simulate ant colonies, but to use artificial ant colonies as an optimization tool. Therefore the method displays several variations in corresponding to the real (natural) ant colony: artificial ants have some memory; they are not blind; they exist in a situation with distinct time. In ACO algorithms, artificial ants create solutions from scratch by probabilistically making a series of local decisions. At each construction stage, an ant chooses precisely one of the possibly numerous ways of stretching the current partial solution. The rules that specify the solution construction method in ACO completely map the search space of the analyzed problem (including the partial solutions) onto a search tree.

The model has established on the observation made by ethologists about the medium used by ants to disclose information concerning shortest paths to food using pheromone trails. A traveling ant lays some pheromone on the ground, thus making a way by a trace of this object. While an isolated ant moves nearly at random (exploration), an ant encountering a previously laid trail can detect it and decide with high probability to follow it and consequently reinforce the path with its pheromone (exploitation). What emerges is a form of the autocatalytic process through which the further ants follow a trail, the more attractive that trail becomes has to be developed. The method has described by a positive feedback loop, during which the probability of determining a path minimizes the number of ants that previously chose the same way. The mechanism above is the inspiration for the algorithms of the ACO family.

4.7 Proposed Entropy-based Fuzzy TOPSIS Decision Making Model

This fuzzy TOPSIS decision-making algorithm is presented. This algorithm is utilized to rank the feature selection methods and to evaluate the process for classification of skin disease. The following steps represent the Intuitionistic Fuzzy TOPSIS decision-making model.

Step 1: Building an intuitionistic fuzzy decision matrix.

Step 2: Determine the criteria weights using the entropy-based method.

Step 3: Create the weighted intuitionistic fuzzy decision matrix.

Step 4: Determine intuitionistic fuzzy positive ideal solution and intuitionistic fuzzy negative ideal solution.

Step 5: Calculate the distance measure of each alternative from Fuzzy positive ideal solution and fuzzy negative ideal solution.

Step 6: Estimate the relative closeness coefficient of each alternative and rank the preference order of all alternatives.

5. RESULT AND DISCUSSION

The following parameters are considered to evaluate the feature selection methods. In the ideal situation, some parameters like accuracy, the true positive rate should have maximum values while others like the number of features, error, should have the least amount. All settings are considered equivalently relevant, and unit weight has allotted to each of them. However, in exceptional circumstances, some parameters may have more effect than the others, so weight has to conform accordingly. The disease survival dataset is considered in this work. This dataset is taken from UCI repository [12].

Table 1: Performance Metrics for Feature selection and Classification

Sl.No	Parameter Name	Desired Values
1	Accuracy	Maximum
2	Root Mean Squared Error	Minimum
3	True Positive Rate	Maximum
4	False Positive Rate	Minimum
5	Precision	Maximum
6	F-Measure	Maximum
7	Receiver Operating Characteristic	Maximum

The classification techniques like Artificial Neural Network, Naïve Bayes Classification, and Support Vector Machine are used in this paper to evaluate the feature selection techniques. The feature selection methods like Gain Ratio (GR), Information Gain (IG), Chi-Square (CS), Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) Feature Selection methods. The following table 2 presented the performance analysis of the existing feature selection methods has estimated by using Support Vector Machine (SVM) as a classification technique.

Table 2: Performance analysis of the original dataset, feature selection methods like GR, GA, IG, CS, PSO and ACO with SVM Classification method

Performance Metrics	Original Dataset	Feature Selection Methods					
		GR	*GA*	*IG*	*CS*	*PSO*	*ACO*
Accuracy (in %)	51.667	73.333	85.333	73.667	63.333	53.667	52.759
RMSE	0.1892	0.5164	0.383	0.5132	0.3291	0.9815	0.8726
TP Rate	0.517	0.733	0.853	0.737	0.133	0.037	0.182
FP Rate	0.741	0.738	0.833	0.737	0.149	0.962	0.811
Precision	0.488	0.542	0.793	0.543	0.128	0.074	0.193
F-Measure	0.502	0.623	0.817	0.625	0.112	0.044	0.156
ROC Area	0.388	0.498	0.51	0.5	0.492	0.037	0.159

Table 3 gives the performance analysis of the original dataset, feature selection methods with Naïve Bayes Classification technique. Table 4 gives the performance analysis of the original dataset, feature selection methods with Support Vector Machine Classification technique.

Table 3: Performance analysis of the original dataset, feature selection methods like GR, GA, IG, CS, PSO and ACO with NB Classification method

Performance Metrics	Original Dataset	Feature Selection Methods					
		GR	*GA*	*IG*	*CS*	*PSO*	*ACO*
Accuracy (in %)	49.333	55.333	51.333	50.000	45.667	52.667	51.448
RMSE	0.6301	0.544	0.6032	0.6	0.6381	0.6082	0.6152
TP Rate	0.493	0.553	0.513	0.5	0.457	0.52	0.4982
FP Rate	0.396	0.481	0.44	0.439	0.564	0.466	0.455
Precision	0.806	0.798	0.8	0.637	0.76	0.84	0.79
F-Measure	0.574	0.629	0.594	0.524	0.544	0.616	0.621
ROC Area	0.554	0.581	0.554	0.544	0.467	0.512	0.536

Table 4: Performance analysis of the original dataset, feature selection methods like GR, GA, IG, CS, PSO and ACO with ANN Classification method

Performance Metrics	Original Dataset	Feature Selection Methods					
		GR	*GA*	*IG*	*CS*	*PSO*	*ACO*
Accuracy (in %)	73.667	87.333	86.667	69.6667	73.667	66.333	61.223
RMSE	0.1845	0.3559	0.3651	0.2508	0.1845	0.5697	0.4786
TP Rate	0.437	0.873	0.867	0.697	0.737	0.663	0.681

FP Rate	0.737	0.854	0.855	0.702	0.69	0.579	0.587
Precision	0.543	0.812	0.794	0.608	0.552	0.567	0.581
F-Measure	0.625	0.823	0.82	0.632	0.631	0.612	0.591
ROC Area	0.442	0.51	0.506	0.497	0.464	0.492	0.488

Table 5 depicts the confidence level of the original dataset, Chi-Squared analysis, Gain Ratio, Information Gain, Genetic Algorithm, Particle Swarm Optimization, and Ant Colony feature selection methods using various classification methods like Naïve Bayes, Artificial Neural Network and Support Vector Machine by using Entropy based TOPSIS multi criteria decision making technique. The confidence value is high for GR analysis, GA and IG feature selection methods than the existing methods using different classification techniques.

Table 5: Confidence level of the original dataset, Chi-Squared analysis, Gain Ratio, Information Gain, GA, PSO and ACO feature selection methods by using NB, ANN and SVM classifiers

Feature Selection Methods	**Classification Techniques**		
	Naïve Bayes	Artificial Neural Network	Support Vector Machine
Original Dataset	0.1577	0.1483	0.1986
GR	0.8587	0.9621	0.7891
GA	0.9216	0.8034	0.7104
IG	0.8011	0.8314	0.8959
CS	0.4462	0.6117	0.2959
PSO	0.2575	0.3226	0.4161
ACO	0.3214	0.4288	0.3643

Table 6 represents the ranking of the feature selection methods using different classification techniques like NB, ANN and SVM.

Table 6: Fuzzy TOPSIS based Ranking of the original dataset, GR, GA, IG, CS, PSO and ACO feature selection methods by using NB, ANN and SVM classifiers

Feature Selection Methods	**Classification Techniques**			**Final Rank**
	Naïve Bayes	Artificial Neural Network	Support Vector Machine	
Original Dataset	7	7	7	**7**
GR	2	1	1	**1**
GA	3	2	2	**2**
IG	1	3	3	**3**
CS	5	5	4	**5**
PSO	4	4	5	**4**
ACO	6	6	6	**6**

6. CONCLUSION

Alternatives can be performed quickly when there is only one feature, but selection among many possible methods sometimes become difficult job because they have multiple metrics. The strength of feature selection methods mostly depends on the variety of dataset. One technique may give excellent effect on one type of dataset but may underperform on a different kind of dataset. TOPSIS can be utilized to recommend one, among some possible alternatives where each choice has several features. In this paper, Entropy-based TOPSIS is used to rank the feature selection techniques like GR, GA, IG, CS, PSO, and ACO based Feature Selection methods. The feature selection techniques have estimated by classification methods. From the above result and discussion, it has confirmed that GR, GA and IG based feature selection methods performs well in the ANN classification technique for improving the prediction of disease survival.

REFERENCE

[1] Prasad, KalliSrinivasaNageswara, and Shameena Begum. "Enhancing Healthcare Systems Using Advanced Data Analytics Techniques."

[2] Nguyen, Sau Huu, et al. "Health-Related quality of life impairment among patients with different skin diseases in Vietnam: a cross-sectional study." *International journal of environmental research and public health* 16.3 (2019): 305.

[3] Liang, Huiying, et al. "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence." *Nature medicine* 25.3 (2019): 433-438.

[4] Wu, Chieh-Chen, et al. "Prediction of fatty liver disease using machine learning algorithms." *Computer methods and programs in biomedicine* 170 (2019): 23-29.

[5] Jo, Taeho, KwangsikNho, and Andrew J. Saykin. "Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data." *Frontiers in aging neuroscience* 11 (2019): 220.

[6] Ngiam, Kee Yuan, and Wei Khor. "Big data and machine learning algorithms for health-care delivery." *The Lancet Oncology* 20.5 (2019): e262-e273.

[7] Kawakami, Eiryo, et al. "Application of artificial intelligence for preoperative diagnostic and prognostic prediction in epithelial ovarian cancer based on blood biomarkers." *Clinical Cancer Research* 25.10 (2019): 3006-3015.

[8] Abdar, Moloud, et al. "A new machine learning technique for an accurate diagnosis of coronary artery disease." *Computer methods and programs in biomedicine* 179 (2019): 104992.

[9] Durairaj, M., and T. S. Poornappriya. "Why Feature Selection in Data Mining Is Prominent? A Survey." *International Conference on Artificial Intelligence, Smart Grid and Smart City Applications*. Springer, Cham, 2019.

[10] Wei, Bo, et al. "Efficient feature selection algorithm based on particle swarm optimization with learning memory." *IEEE Access* 7 (2019): 166066-166078.

[11] Fu, Mo. "A Heuristic Search Method Based on GA-SA Algorithm." *International Conference on Applications and Techniques in Cyber Security and Intelligence*. Springer, Cham, 2019.

[12] https://archive.ics.uci.edu/ml/datasets/HCC+Survival