

An Adaptive Soft Set Based Diagnostic Risk Prediction System

Terry Jacob Mathew^{1(✉)}, Elizabeth Sherly², and José Carlos R. Alcantud³

¹ School of Computer Sciences, Mahatma Gandhi University, Kottayam, India
terryjacobin@gmail.com

² IIITM-K, Technopark, Thiruvananthapuram, India
sherly@iiitmk.ac.in

³ BORDA Research Unit and Multidisciplinary Institute of Enterprise (IME),
University of Salamanca, 37007 Salamanca, Spain
jcr@usal.es
<http://diarium.usal.es/jcr>

Abstract. Recently, risk based prediction models in medical diagnostic systems gain wider significance in deciding most appropriate diagnostic treatments and for clinical usage. Prostate cancer is a disease which is difficult to diagnose and there are number of failure cases reported. Therefore, an effective and aggressive selection of multiple factors influence on the disease is required. In this paper, an adaptive soft set based diagnostic risk prediction system is presented with the implementation on prostate cancer. The system receives input parameters related to the disease and gives out the risk percentage of the patient. Soft sets are generated with the input parameters by fuzzification followed by rule generation. The risk percentage of the rules are individually calculated for Precision, Recall and F-Measure, that conclude on the best risk percentage based on the maximum area under the curve (AUC) in each case. This ensures to select the most influential risk parameters in treating the disease. Specificity and sensitivity of the test system yield 75.00% and 45.45% respectively.

Keywords: Soft sets · Fuzzy set · Prostate cancer · Decision making

1 Introduction

The presence of intelligent systems in the field of medical sciences have been undergoing phenomenal growth for the last two decades. Earlier, expert systems have significantly influenced the way a doctor deals with and diagnose a patient. Some notable examples are MYCIN [33], INTERNIST [25], etc. Further more, the advances in information technology embraces the digitization of the medical records and the development of other technology related medical applications in an accelerated pace.

The use of applications involving artificial intelligence and machine learning can successfully assist physicians with distinctive diagnosis of diseases, treatment opinions and recommendations, radio diagnosis on images etc. Data mining in health care also provide “real time” diagnostic as well as effective medicine recommendations on the basis of a training data set. But, driven by the issues in existing risk prediction systems, an adaptive method to improve the scope and accuracy of the prediction system is presented.

The organization of this paper is as follows. Section 2 discusses the related works. Section 3 gives an overview of the adaptive soft set based risk prediction system along with the preliminaries in Subsect. 3.1. The proposed algorithm is given in Subsect. 3.2. The implementation details of the proposed algorithm for prostate cancer detection is given in Sect. 4 along with the detailed description of each step. Section 5 discusses the results and implementation details of the algorithm. We conclude in Sect. 6.

2 Related Work

Fuzzy sets and fuzzy logic were introduced by Zadeh [38] to handle problems with uncertainty, and since then, they have contributed to a paradigm shift in the way we deal with imprecise problems and their solutions. Fuzzy sets are highly successful in many areas and give improved results than what the classical approach does. However, fuzzy set depends on membership function to represent impreciseness and are subjective to the user-level intervention. This often degrades the overall performance. Hence, to solve these problems, many researchers put forward solutions, like Atanassov [5] put forward the concept of intuitionistic fuzzy sets; Pawlak [29] introduced rough sets; Torra [35] put forward the hesitant fuzzy sets; which are applicable in real-world situations as in Alcantud et al. [1], etc. Nevertheless, these theories are limited due to the lack of parameterization concept associated with them for describing the problem.

In 1999, Molodtsov [26] introduced soft sets and established the fundamental results of this theory. A soft set is a collection of approximate descriptions of an object and is used as a general mathematical tool for dealing with objects which have been defined using a very loose and hence very general set of characteristics. Molodtsov further showed that soft set theory is free from parameterization inadequacy syndrome. Ali et al. [4] and Feng and Li [18] also contributed to settle the fundamental laws that govern this notion. Extensions and hybrid models that combine the soft set model with others have been defined and used for decision making e.g., in Ali [3], Das [11], Feng [16], Feng et al. [17, 19], Ma et al. [22], Peng and Yang [30] Zhan et al. [39] and [15].

Recently the use of soft set based intelligent systems have gained interest in computational intelligence by giving better solutions for compound problems. Soft set theory and fuzzy set theory have been successfully used in some medical systems, for example [9, 27, 28, 36]. The intelligent systems in medicine based on soft sets, generally depend on finding the risk percentage as a single value and then use it to grade the severity of the condition. This unilateral approach will not give a comprehensive picture of the real risk percentage of the patient.

We take up receiver operator characteristic (ROC) curve as a preferred performance evaluation tool to validate classifier performance over a range of decision thresholds [10,13]. The area under the curve (AUC), has been traditionally used in medical diagnosis since the 1970's [20]. The AUC maps the entire ROC curve into single number, that reflects the overall performance of the classifier over all thresholds.

Prostate cancer is the common cause of cancer death among men and it depends on various elements such as hereditary factors, age, ethnic background, the level of prostate specific antigen (PSA) in the blood etc. The level of PSA in the blood is a very important indicator for an initial diagnosis in patients [7]. But, as multiple factors cause the level of PSA to fluctuate, there exists uncertainty in the diagnosis. A biopsy of the prostate can give a distinctive diagnosis of cancer. But all patients will not be cancer positive after the biopsy. Also, it is always better to avoid an unnecessary biopsy as doctors and researchers have noted that biopsy of a tumour can cause spread of cancer cells leading to multiple sites of tumour at the biopsy site [14]. To help the doctor detect the patients with low risk of prostate cancer, a decisive intelligent system with more significant risk percentage prediction is needed.

2.1 Existing Methods and the Scope for a New Proposal

Soft computing is a host of methodologies which work in unison for providing flexible information processing capability for handling real life uncertain situations. It exploits the tolerance for imprecision, uncertainty, approximate reasoning and partial truth to build low-cost solutions. Apart from fuzzy logic, neural networks, and genetic algorithm methodologies, emergence of soft set based medical prediction systems are also on the rise. We take a few for evaluation.

Sanchez [31] initiated the use of fuzzy techniques to possibility distributions in natural languages and medical diagnosis. This notion was later extended with intuitionistic fuzzy sets by De et al. [12]. Slowiński [34] experimented with 122 patients treated for duodenal ulcer by applying rough sets to create a decision algorithm which could be used in the treatment of new ulcer patients. A pioneering prediction system for calculating the patient's prostate cancer risk given in [36]. Yuksel et al. [37] combined covering soft set and rough sets to produce soft covering based rough set and applied it for prostate cancer diagnosis. In another method given by Feng in [16], soft covering based rough sets are applied to a medical problem of calculating the risk of prostate cancer. Some other papers which deal with prostate cancer risk prediction are [6,21]. A recent contribution for glaucoma detection is given in [2]. In all the above mentioned papers, the risk percentage is calculated by a single metric based on their respective methods.

The main purpose of this paper is to provide a specific approach to improve the soft set based prediction system. Our adaptive prediction system model is tested with the data of 120 patients with prostate complaint from Selcuk University Meram Medicine Faculty [32]. We are interested in improving the accuracy of the predicted risk percentage by including the prudent use of metrics like Precision, Recall and F-measure. The Precision results depend only on the retrieved

result subsets of the actual data. It does not consider the total positives in the whole of the data. If we consider Recall for analysing the risk percentage, then the total patients are considered for risk calculation. But, the issue with Recall is that false positives cannot be discerned. To compensate for the drawbacks of Precision and Recall, we depend on F-measure. The traditional F-measure is the harmonic mean of Precision and Recall.

We propose to include the tradeoff between these metrics into the risk prediction process. In [36], the risk is calculated as follows. For example, if there are 13 patients satisfying say Rule 1, of which 4 are found to be prostate cancer positive, then the risk percentage for Rule 1 is calculated as $(4 \div 13) \times 100 = 30.76$. The patients are compared with compatible rules and the highest risk percentage is accorded as the risk percentage of the patients. Here, the drawback of using Precision can be explained by this assumption. Assume that out of the compatible rules generated for a patient; say Rule 1 has only 1 patient and this patient is also tested positive in biopsy, then the patient will be awarded with 100% risk percentage. This may not be always true.

As there is no qualitative and quantitative analysis on the result output, these single handed general methods of deriving the risk percentage from soft set based and other prediction techniques are not a legitimate way of calculation. Therefore, we propose to include Precision, Recall and F-measure in finding the significant associated risk. Also, rather than taking the highest observed risk percentage of a specific rule as the risk of the patient, an average of the most relevant risks are calculated. This averaging of the selected rules make our model more effective.

3 Adaptive Soft Set Based Risk Prediction System

In this section, we present the basic definitions of soft set theory [26] and fuzzy set theory [23] followed by the proposed system. These definitions and further details on soft sets and fuzzy sets can be found in [8, 24, 38]. As usual, we follow the common terminology for describing soft set and its extensions. Here U refers to an initial universe and E is the set of parameters.

3.1 Definitions: Soft Set and Fuzzy Set

Definition 1 (Molodtsov [26]). A pair (F, A) is a soft set over U when $A \subseteq E$ and $F : A \rightarrow \mathcal{P}(U)$, where $\mathcal{P}(U)$ denotes the set of all subsets of U .

Example 1. A soft set over U is regarded as a parameterized family of subsets of the universe U , the set A being the parameters. For each parameter $e \in A$, $F(e)$ is the subset of U approximated by e or the set of e -approximate elements of the soft set.

Let U be the set of five patients given by $U = \{p_1, p_2, p_3, p_4, p_5\}$ and E be the set of symptoms given by $E = \{s_1, s_2, s_3, s_4, s_5\}$.

Let $A = \{s_1, s_2, s_3\}$ be the set of symptoms, the doctor intends to use for diagnosis. Now consider that (F, A) is a mapping given by, $(F, A)(s_1) = \{p_1, p_2\}$, $(F, A)(s_2) = \{p_1, p_3\}$ and $(F, A)(s_3) = \{p_2, p_4\}$.

Then the soft set $(F, E) = \{(s_1, \{p_1, p_2\}), (s_2, \{p_1, p_3\}), (s_3, \{p_2, p_4\}), (s_4, \{\emptyset\}), (s_5, \{\emptyset\})\}$. A soft set can also be represented in the form of a two dimensional table. Table 1 is the tabular representation of the soft set (F, E) shown in Example 1.

Table 1. Tabular representation of the fuzzy set (F, A) associated with Example 1.

| U/E | s_1 | s_2 | s_3 | s_4 | s_5 |
|-------|-------|-------|-------|-------|-------|
| p_1 | 1 | 1 | 0 | 0 | 0 |
| p_2 | 1 | 0 | 1 | 0 | 0 |
| p_3 | 0 | 1 | 0 | 0 | 0 |
| p_4 | 0 | 0 | 1 | 0 | 0 |
| p_5 | 0 | 0 | 0 | 0 | 0 |

Definition 2 (Maji et al. [24]). Let (F, A) and (G, B) be two soft sets. Then the AND operation of (F, A) AND (G, B) , denoted by $(F, A) \wedge (G, B)$, is defined as $(H, A \times B)$ where $H(\alpha, \beta) = F(\alpha) \cap G(\beta)$ for each $(\alpha, \beta) \in A \times B$.

Definition 3 (Zadeh [38]). A fuzzy set X over U is a set defined by a function μ_X representing a mapping $\mu_X : U \rightarrow [0, 1]$. where, μ_X is called the membership function of X , and the value $\mu_X(u)$ is the grade of membership of $u \in U$. The value of $\mu_X(u)$ represents the degree with which u belongs to the fuzzy set X . Thus, a fuzzy set X over U can be represented as follows:

$$X = \{(\mu_X(u)/u) : u \in U, \mu_X(x) \in [0, 1]\}.$$

For a fuzzy set X in U and any real number $\alpha \in [0, 1]$, then the α -cut or cut worthy set of A , denoted by $X[\alpha]$ is the crisp set defined as $\{x \in U : \mu_X(x) \geq \alpha\}$.

3.2 Proposed Intelligent System for Prostate Cancer Diagnosis

The available data set is attributed with a set of three variables, namely prostate specific antigen (PSA), prostate volume (PV) and age of the patient. The membership function of these variables are shown in Eqs. (1) and (2). All 120 selected patients underwent biopsy and their diagnostic results are known. In the following part, we proceed to explain the step by step procedures, which make up the proposed algorithm.

In order to facilitate the representation of soft sets, we initially convert the input data into fuzzy sets. Afterwards, going by the principle of including fuzzy

sets as soft sets (cf., Molodtsov [26]), the fuzzy sets are redeployed correspondingly as relevant soft sets. Unlike the conventional soft set prediction methods, we avoided parameter reduction in view of the nature and type of data set. The decisive phase is generation of rules, which are analysed later for determining the prostate cancer risk. Each rule is awarded a risk percentage which determines the verdict of the intelligent system. The algorithm for prostate cancer detection with stepwise descriptions is given below.

AdaptiVe Algorithm for Softset based predicTion (AVAST).

- Step 1.** Fuzzyfication of data set with the selected variables namely PSA, PV and age.
- Step 2.** Transforming the fuzzy sets corresponding to input data into soft sets.
- Step 3.** Obtaining the rules relevant for the system by the application of AND operator on to the soft sets generated in the previous step.
- Step 4.** Analysis of rules based on Precision, Recall and F-measure.
- Step 5.** Plot the ROC curve with the calculated risk percentage for the above three sets.
- Step 6.** Select the metric i.e. (either Precision, Recall or F-measure), which offers the maximum AUC and proceed for actual risk prediction over the testing set.

4 Implementation of Algorithm - AVAST to Calculate Prostate Cancer Risk

The various stages of algorithm - AVAST is explained in detail below.

Explanation of Step 1. We fuzzificate the patient data with appropriate membership functions on the basis of inputs from medical literature [36]. The following linguistic variables are modelled for the attributes PSA, PV and age. The PSA variables VL, L, M, H and VH represent very low, low, middle, high and very high respectively. The PV variables S, M, B and VB represent small, medium, big and very big respectively. The age factor attributed by VY, Y, M and O represents very young, young, middle and old respectively. Trapezoidal or triangular membership functions can be selected for each variable on the basis of their interval size. The corresponding membership values are determined from Eqs. 1 and 2.

$$PSA(x) = \begin{cases} \mu_x & \text{if } 0 < x < 100 \\ 1 & \text{if } x \geq 100 \end{cases} \quad PV(y) = \begin{cases} \mu_y & \text{if } 30 < y < 120 \\ 1 & \text{if } y \geq 120 \end{cases} \quad (1)$$

$$Age(z) = \begin{cases} 0 & \text{if } z \leq 20 \\ \mu_z & \text{if } 20 < z < 65 \\ 1 & \text{if } z \geq 65 \end{cases} \quad (2)$$

Table 2. A sample input data of patients

| U | Age | PSA | PV |
|----------|-------|-------|------|
| U_7 | 54 | 5.62 | 28 |
| U_9 | 54 | 17.3 | 45 |
| U_{20} | 59 | 8.36 | 55 |
| U_{34} | 61 | 18.3 | 62 |
| U_{40} | 62 | 51.74 | 29 |
| U_{70} | 68 | 140 | 117 |
| U_{99} | 73 | 47.4 | 87 |

A sample of the input data is shown in Table 2 and the parameter memberships are shown in Fig. 1.

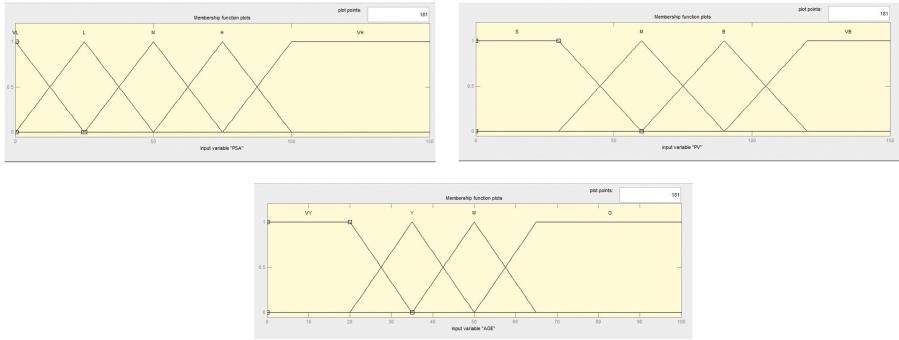


Fig. 1. The membership functions of Age, PSA and PV

Explanation of Step 2. We directly depend on Molodtsov’s method for the transformation of fuzzy sets into soft sets. The Molodtsov’s method maps soft sets on to the universe $[0, 1]$, thus making the selection of a subset of this range inevitable to conduct a practical setting of this experiment. Depending on the distribution of patient data into different levels of membership, we have different elements in different subsets for each variables. Table 3 shows the fuzzy membership values of the input factors. In this approach, for a soft set (F, A) over U , A is denoted by a set of parameters represented by $\{e_1, e_2, e_3, e_4, \dots, e_n\}$, then $F(e_j)$ is a subset of U , where, $F(e_j) = \{U_i \mid \mu(U_i) \geq e_j; \forall j = 1 \text{ to } n \text{ and } i = 1 \text{ to } m\}$.

Hence, the newly formed soft sets will have subsets of elements from the universal set U . As an example from the data set, the soft set,

Table 3. The fuzzy membership values of factors shown in Table 2

| U | Age | PSA | PV |
|----------|----------------|-----------------|----------------|
| U_7 | 0.73 M, 0.26 O | 0.77 VL, 0.22 L | 1 S |
| U_9 | 0.73 M, 0.26 O | 0.30 VL, 0.69 L | 0.5 S, 0.5 M |
| U_{20} | 0.4 M, 0.6 O | 0.66 VL, 0.33 L | 0.16 S, 0.83 M |
| U_{34} | 0.26 M, 0.73 O | 0.26 VL, 0.73 L | 0.93 M, 0.06 B |
| U_{40} | 0.2 M, 0.8 O | 0.93 M, 0.06 H | 1 S |
| U_{70} | 1 O | 1 VH | 0.1 B, 0.9 VB |
| U_{99} | 1 O | 0.10 L, 0.89 M | 0.1 M, 0.9 B |

$F : A_{Age(M)} \longrightarrow \mathcal{P}(U)$ is associated with a parameter set,

$A_{Age(M)} = \{0.06, 0.28, 0.5, 0.71, 0.93\}$, and the corresponding soft sets obtained are as:

$$F(.06) = \{U_{42}, U_{43}, U_{44}, U_{45}, U_{46}, U_{41}, U_{35}, U_{37}, U_{40}, U_{30}, U_{31}, U_{32}, U_{33}, U_{34}, U_{25}, U_{27}, U_{28}, U_{29}, U_{19}, U_{20}, U_{21}, U_{22}, U_{23}, U_{24}, U_{15}, U_{16}, U_{17}, U_{18}, U_1, U_{13}, U_{10}, U_8, U_9, U_4, U_5, U_2, U_3\}$$

$$F(.28) = \{U_{25}, U_{27}, U_{28}, U_{29}, U_{19}, U_{20}, U_{21}, U_{22}, U_{23}, U_{24}, U_{15}, U_{16}, U_{17}, U_{18}, U_1, U_{13}, U_{10}, U_8, U_9, U_4, U_5, U_2, U_3\}$$

$$F(.5) = \{U_1, U_{13}, U_{10}, U_8, U_9, U_4, U_5, U_2, U_3\},$$

$$F(.71) = \{U_8, U_9, U_4, U_5, U_2, U_3\}, \text{ and}$$

$$F(.93) = \{U_2, U_3\}$$

Explanation of Step 3. The combination of soft sets obtained in Step 2 by AND'ing operation gives all possible rules. By this means, a total of 1760 rules are generated, which are checked for compatibility with the patients. An obtained sample rule is given below.

For example: $AGE(M)(.5) \wedge PSA(VL)(.05) \wedge PV(M)(.5) = \{u_{10}, u_{14}\}$.

Explanation of Step 4. The output obtained from Step 3 is processed further to associate each rule with a risk of prostate cancer as follows.

The rules obtained above will have patients from the training data and the Precision, Recall and F-measure based risk percentage is calculated for each rule. The calculated risk percentage of patients are then separately considered and averaged individually. Thus, corresponding to each test data, we now have separate risk percentage for Precision, Recall and F-measure.

For a sample training data, the rules obtained are as:

$$\begin{aligned}
 R_1 &= \{U_{10}, U_{14}\} \\
 R_2 &= \{U_{13}, U_{17}, U_{81}\} \\
 R_3 &= \{U_{21}, U_{22}, U_{54}, U_{67}, U_{98}, U_{107}\} \\
 R_4 &= \{U_3, U_8, U_{21}, U_{22}, U_{54}, U_{67}, U_{98}, U_{99}, U_{107}\} \text{ and} \\
 R_5 &= \{U_{40}, U_{92}, U_{116}\}
 \end{aligned}$$

The biopsy results for the patients are known from the labelled data set and the calculated values of Precision, Recall and F-measure (F1) for the above rules are calculated as per Eq. 3 and are shown in the Table 4. Here TP indicates true positive, FN is false negative and FP is false positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \text{ Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

Some of the validation results of rules' risk percentage based on Precision, Recall and F-measure are shown in Table 5, where test samples with "*" indicate the outliers.

Table 4. The risk percentage of rules on the basis of Precision, Recall and F-Measure

| Rules | Precision | Recall | F-Measure |
|-------|-----------|--------|-----------|
| R_1 | 0 | 0 | 0 |
| R_2 | 0 | 0 | 0 |
| R_3 | 0.5 | 0.06 | 0.11 |
| R_4 | 0.56 | 0.1 | 0.18 |
| R_5 | 1 | 0.06 | 0.12 |

Table 5. The risk percentage of a set of patients by Precision, Recall and F-measure

| Test data | Risk percentage while considering | | | Ground truth |
|------------|-----------------------------------|--------|-----------|--------------|
| | Precision | Recall | F-Measure | |
| U_1 | 38.29 | 10.69 | 7.92 | 0 |
| U_{10} | 38.36 | 13.35 | 8.81 | 0 |
| U_{31} | 43.27 | 17.68 | 11.37 | 0 |
| U_{43} | 47.91 | 28.5 | 16.83 | 0 |
| U_{60} | 54.08 | 30.17 | 18.66 | 1 |
| U_{77*} | 41.52 | 5.68 | 4.4 | 1 |
| U_{83} | 55.35 | 22.35 | 14.49 | 1 |
| U_{86} | 54.35 | 28.57 | 17.84 | 1 |
| U_{95*} | 54.51 | 25.08 | 16.31 | 0 |
| U_{113*} | 55.09 | 21.55 | 14.47 | 0 |

The next step will determine, whether we will consider Precision, Recall or F-measure, for calculating the actual risk percentage for each patient.

Explanation of Step 5. Plot the ROC curve for the risk percentage based on Precision, Recall and F-measure and select the risk percentage for the test patients corresponding to the maximum AUC. The AUC maps the entire ROC curve into single value that portrays the overall performance of the classifier over all thresholds. The false positive rate (FPR) and true positive rate (TPR) evaluate performance for a specific threshold. The FPR and TPR can also be combined to form an overall mis-classification rate, which is known as true error. By getting the ROC, AUC, FPR and TPR of this system, we obtain the complete knowledge about the performance of the system. Table 6 shows the AUC generated for some test groups.

Table 6. The AUC of some samples used in generating the ROC curve

| Test data sets | Area Under the Curve (AUC) | | |
|----------------|----------------------------|--------|-----------|
| | Precision | Recall | F-Measure |
| Dataset 1 | 40.15 | 56.43 | 56.43 |
| Dataset 2 | 60.07 | 58.68 | 59.72 |
| Dataset 3 | 50.69 | 68.75 | 53.13 |
| Dataset 4 | 45.49 | 49.31 | 48.96 |
| Dataset 5 | 63.54 | 46.53 | 44.44 |

By employing a five-fold cross validation, the patient data is randomly divided into training and testing sets. By this approach [13], the original data is randomly partitioned into five equal sized sub-samples. Of the five sub-samples, a single sub-sample is retained as the validation data for testing the model, and the remaining four sub-samples are used as training data. The cross-validation process is then repeated five times (the folds). The five results from the folds can then be averaged to produce a single estimation.

Explanation of Step 6. Finally, the choice of the appropriate metric is done from Precision, Recall and F-measure on the basis of the maximum AUC generated over the validation data.

5 Results and Discussion

After five fold cross validation, the ROC plots corresponding to Precision, Recall and F-Measure for validation data are shown in Fig. 2. The selected metric can be either Precision, Recall or F-Measure based on the AUC obtained for each case. In this investigated case, as seen in Fig. 2, Precision based ROC curve has the highest AUC than Recall and F-Measure. Hence Precision will be selected for calculating the rule risk percentage.

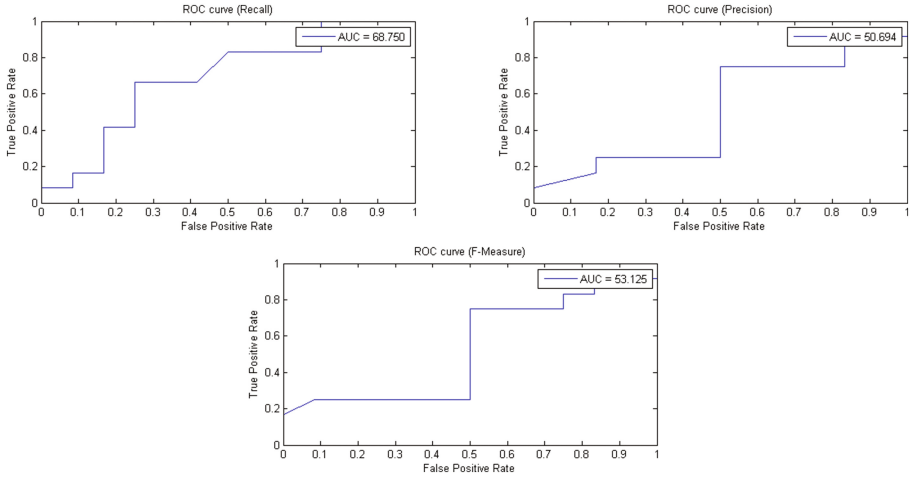


Fig. 2. The ROC curve for Precision, Recall and F-measure

The proposed algorithm model was implemented by means of scientific computation platform R2013a Matlab. In this investigation, we have divided the total data into testing data, training data and validation data. One set is for training, one for testing and the other two for validation. By the application of AVAST algorithm on the validation data, we could select the best from the Precision based, Recall based and F-Measure based methods on the basis of maximum AUC.

As sensitivity is significant for this specific case of prostate cancer prediction, we have to select a threshold value which gives high sensitivity over the validation process. Concurrently the false positive rate (FPR) should be minimal. So a particular risk percentage value is selected as the threshold when the corresponding TPR is at least 80% and with minimum FPR. This threshold is then applied for the final test data. Figure 3 shows the ROC curve for the test system. It should be noted that we can change the threshold selection parameters for

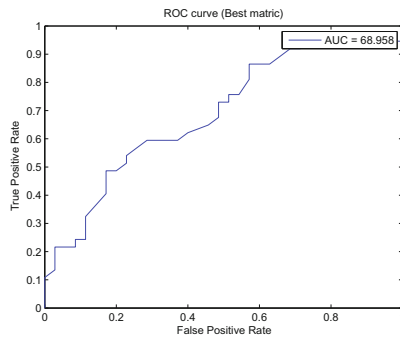


Fig. 3. The ROC curve for the test data

the system on the basis of data set and other requirements. The sensitivity and specificity of the test system stood at 75.00% and 45.45 % respectively.

6 Conclusions

In this work, it is shown that the soft set approach for finding the risk percentage of prostate cancer patients can be significantly improved with the inclusion of Precision, Recall and F-measure based rule analysis. The proposed method exhibits an adaptive nature as the best performing metric is chosen on the basis of the validation data set performance. We depend on these statistical measures to optimize the risk percentage calculation. This general notion can be applied to all methods which follow a unidirectional approach in defining the output risk percentage. The results confirm that the inclusion of this adaptive approach to existing methodologies show better results as shown in Sect. 4.

As future enhancements, we propose to extend our approach with other existing algorithms with more relevant parameters for reducing ambiguity. Also a weighted approach for rules and parameters can be employed to see if it leads to further improvements. The medical applications using the concepts of soft sets opens up lot of room for exploration and innovation. A quick and automated method of prostate cancer diagnosis based on soft sets is addressed in this contribution. The proposed model helps the doctor to discern the patients for the biopsy procedure to detect prostate cancer.

References

1. Alcantud, J.C.R., de Andres Calle, R., Torrecillas, M.J.M.: Hesitant fuzzy worth: an innovative ranking methodology for hesitant fuzzy subsets. *Appl. Soft Comput.* **38**, 232–243 (2016)
2. Alcantud, J.C.R., Santos-García, G., Hernández-Galilea, E.: Glaucoma diagnosis: a soft set based decision making procedure. In: *Conference of the Spanish Association for Artificial Intelligence*, pp. 49–60. Springer (2015)
3. Ali, M.: A note on soft sets, rough soft sets and fuzzy soft sets. *Appl. Soft Comput.* **11**, 3329–3332 (2011)
4. Ali, M.I., Feng, F., Liu, X., Min, W.K., Shabir, M.: On some new operations in soft set theory. *Comput. Math. Appl.* **57**(9), 1547–1553 (2009)
5. Atanassov, K.: Intuitionistic fuzzy sets. *Fuzzy Sets Syst.* **20**, 87–96 (1986)
6. Benecchi, L.: Neuro-fuzzy system for prostate cancer diagnosis. *Urology* **68**(2), 357–361 (2006)
7. Catalona, W.J., Partin, A.W., Slawin, K.M., Brawer, M.K., Flanigan, R.C., Patel, A., Richie, J.P., Walsh, P.C., Scardino, P.T., Lange, P.H., et al.: Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease: a prospective multicenter clinical trial. *Jama* **279**(19), 1542–1547 (1998)
8. Çağman, N., Enginoğlu, S.: Soft set theory and uni-int decision making. *Eur. J. Oper. Res.* **207**(2), 848–855 (2010)
9. Çelik, Y., Yamak, S.: Fuzzy soft set theory applied to medical diagnosis using fuzzy arithmetic operations. *J. Inequalities Appl.* **2013**(1), 82 (2013)

10. Cohn, T.E.: Receiver operating characteristic analysis of photoreceptor sensitivity. *IEEE Trans. Syst. Man Cybern.* **5**, 873–881 (1983)
11. Das, A.K.: Weighted fuzzy soft multiset and decision-making. *Int. J. Mach. Learn. Cybern.* 1–8 (2016). Springer
12. De, S.K., Biswas, R., Roy, A.R.: An application of intuitionistic fuzzy sets in medical diagnosis. *Fuzzy Sets Syst.* **117**(2), 209–213 (2001)
13. D’Errico, G.E.: Receiver operating characteristic: a tool for cell confluence estimation. In: 2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 576–579. IEEE (2015)
14. Eriksson, M., Reichardt, P., Hall, K.S., Schütte, J., Cameron, S., Hohenberger, P., Bauer, S., Leinonen, M., Reichardt, A., Davis, M.R., et al.: Needle biopsy through the abdominal wall for the diagnosis of gastrointestinal stromal tumour-does it increase the risk for tumour cell seeding and recurrence? *Eur. J. Cancer* **59**, 128–133 (2016)
15. Fatimah, F., Rosadi, D., Hakim, R.F., Alcantud, J.C.R.: Probabilistic soft sets and dual probabilistic soft sets in decision-making. In: *Neural Computing and Applications*, pp. 1–11 (2017)
16. Feng, F.: Soft rough sets applied to multicriteria group decision making. *Ann. Fuzzy Math. Inform.* **2**(1), 69–80 (2011)
17. Feng, F., Li, C., Davvaz, B., Ali, M.: Soft sets combined with fuzzy sets and rough sets: a tentative approach. *Soft Comput.* **14**(9), 899–911 (2010)
18. Feng, F., Li, Y.: Soft subsets and soft product operations. *Inf. Sci.* **232**, 44–57 (2013)
19. Feng, F., Liu, X., Leoreanu-Fotea, V., Jun, Y.B.: Soft sets and soft rough sets. *Inf. Sci.* **181**(6), 1125–1137 (2011)
20. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng.* **17**(3), 299–310 (2005)
21. Keles, A., Hasiloglu, A.S., Keles, A., Aksoy, Y.: Neuro-fuzzy classification of prostate cancer using NEFCLASS-J. *Comput. Biol. Med.* **37**(11), 1617–1628 (2007)
22. Ma, X., Liu, Q., Zhan, J.: A survey of decision making methods based on certain hybrid soft set models. *Artif. Intell. Rev.* **47**(4), 507–530 (2017)
23. Maji, P., Biswas, R., Roy, A.: Fuzzy soft sets. *J. Fuzzy Math.* **9**, 589–602 (2001)
24. Maji, P., Biswas, R., Roy, A.: Soft set theory. *Comput. Math. Appl.* **45**, 555–562 (2003)
25. Miller, R.A., Pople Jr., H.E., Myers, J.D.: Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. *New Engl. J. Med.* **307**(8), 468–476 (1982)
26. Molodtsov, D.: Soft set theory - first results. *Comput. Math. Appl.* **37**, 19–31 (1999)
27. Oniško, A., Druzdzel, M.J.: Impact of precision of bayesian network parameters on accuracy of medical diagnostic systems. *Artif. Intell. Med.* **57**(3), 197–206 (2013)
28. Park, K.S., Chae, Y.M., Park, M.: Developing a knowledge-based system to automate the diagnosis of allergic rhinitis. *Biomed. Fuzzy Hum. Sci. Official J. Biomed. Fuzzy Syst. Assoc.* **2**(1), 9–18 (1996)
29. Pawlak, Z.: Rough sets. *Int. J. Comput. Inf. Sci.* **11**(5), 341–356 (1982)
30. Peng, X., Yang, Y.: Algorithms for interval-valued fuzzy soft sets in stochastic multi-criteria decision making based on regret theory and prospect theory with combined weight. *Appl. Soft Comput.* **54**, 415–430 (2017)
31. Sanchez, E.: Inverses of fuzzy relations. Application to possibility distributions and medical diagnosis. *Fuzzy Sets Syst.* **2**(1), 75–86 (1979)

32. Saritas, I., Allahverdi, N., Sert, I.U.: A fuzzy approach for determination of prostate cancer. *Int. J. Intell. Syst. Appl. Eng.* **1**(1), 1–7 (2013)
33. Shortliffe, E.: *Computer-Based Medical Consultations: MYCIN*, vol. 2. Elsevier, New York (2012)
34. Slowinski, K.: Rough classification of HSV patients. In: *Intelligent Decision Support-Handbook of Applications and Advances of the Rough Sets Theory*, pp. 77–94 (1992)
35. Torra, V.: Hesitant fuzzy sets. *Int. J. Intell. Syst.* **25**(6), 529–539 (2010)
36. Yuksel, S., Dizman, T., Yildizdan, G., Sert, U.: Application of soft sets to diagnose the prostate cancer risk. *J. Inequalities Appl.* **2013**(1), 229 (2013)
37. Yüksel, Ş., Tozlu, N., Dizman, T.H.: An application of multicriteria group decision making by soft covering based rough sets. *Filomat* **29**(1), 209–219 (2015)
38. Zadeh, L.: Fuzzy sets. *Inf. Control* **8**, 338–353 (1965)
39. Zhan, J., Liu, Q., Herawan, T.: A novel soft rough set: soft rough hemirings and corresponding multicriteria group decision making. *Appl. Soft Comput.* **54**, 393–402 (2017)