International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 03 | Mar-2016 www.irjet.net p-ISSN: 2395-0072

PREDICTING SURVIVAL OF BREAST CANCER PATIENTS USING FUZZY RULE BASED SYSTEM

Nivetha S .[1], Samundeeswari E.S.[2]

[1] *Research Scholar, Department of Computer Science, Vellalar College for Women Tamilnadu, India*
[2] *Associate Professor, Department of Computer Science, Vellalar College for Women Tamilnadu, India*

Abstract - ***One of the most common techniques for predicting breast cancer proposed in literature is supervised ML methods and classification algorithms. A semi supervised ML techniques can be a great alternative to the other two types of ML. A small sized training sample that compared to data dimensionality can result in misclassifications as the estimators may produce unstable and biased models. Apart from the data size, the dataset quality as well as the careful feature selection schemes is of great importance for effective ML and subsequently for accurate cancer predictions.***

For Breast cancer survival prediction, information like histological and clinical data along with genomic and proteomic information about a Breast cancer patient is required.To effectively predict the survivability of cancer, the integration of multidimensional heterogeneous data, different techniques for feature selection and classification are required.

A variety of techniques have been widely applied in cancer research for the development of predictive models resulting in effective and accurate decision making. The use of fuzzy methodologies has become more important in addressing classification problems over recent years. Specifically, fuzzy rule-based systems have been utilized to produce high classification accuracy through philological rule sets. The heterogeneous features are quantified by using fuzzy subsethood. The fuzzy subset hood is an efficient feature selection method for heterogeneous dataset. In this work, SVM and fuzzy based classifications are used for predicting and the results are validated and compared. It is found that fuzzy based classification yields better results than SVM model.

Key Words: Breast Cancer, Classification, Preprocessing, SVM, Fuzzy method.

1. DATA MINING

Data mining is the process of mining the hidden patterns from huge data. Data mining scans a huge volume of data to find out the patterns and correlations among the patterns. Data mining requires the use of data analysis tool containing statistical model, mathematical algorithms and machine learning methods to determine previously unknown, valid patterns and relationships in huge volume data. Thus, data mining consists of more than gathering and running data, it also contains analysis and prediction.

Data mining can be executed on data in quantitative, textual or multimedia forms. It contains tools such as association, sequence analysis, classification, clustering and forecasting. Data Mining is mostly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.

1.1 SURVIVABILITY

Survivability is the capability to remain alive or continue to exist. Survival analysis is the branch of statistics that involves the modeling of time to event data; in this context, death or failure is considered an "event" in the survival analysis literature.Traditionally, only a single event occurs for each subject, after which the organism or mechanism is dead or broken. The recurring event or repeated event models relax that assumption. Cancer survival is commonly higher in people diagnosed aged under 40 years old, with the exception of breast, bowel and prostate cancers, where survival is highest in middle age. Survival statistics give an overall picture of survival in all adults (15-99) diagnosed stages and co-morbidities. The survival time experienced by an individual patient may be much higher or lower and is based on the specific patient and tumour characteristics. The survival rate refers to the percentage of patients who live at least 5 years after being diagnosed with cancer. Many of these patients live much longer than 5 years after diagnosis. The relative survival rate compares to the observed survival helps to correct the death caused by something besides cancer and is a more accurate way to describe the effect of cancer.

2. ANALYSIS OF RELATED WORK

To better understand survival prediction, it is necessary to review and examine the existing research works**. Mei-Yin C. Polley et al [5]** has discussed about the biomarkers to guide the therapy for patient's disease as informed by biological characterization. For patients with cancer, these characterizations are typically achieved by molecular analysis of tumor biomarkers or by examination of host characteristics, such as variations in germline DNA. Two classes of biomarkers in oncology are prognostic markers and predictive markers. The prognostic markers inform about likely disease outcome determined after the treatment is received, and predictive markers provide information about likely outcomes with application of particular interventions. Therefore, predictive markers can help select among two or more therapy options. Predictive markers are substance for targeted therapies, which often are expected to benefit only patients whose disease is characterized by presence of a biomarker.

Predictive biomarkers guide therapy for cancer patients are a cornerstone of precision medicine. Specific issues addressed are differentiation between qualitative and quantitative predictive effects and challenges due to sample size requirements for predictive biomarker assessment, and consideration of additional factors that relevant to clinical utility assessment, such as toxicity and cost of new therapies as well as costs and potential morbidities associated with the routine use of biomarker-based tests.

Dharanija Madhavan et al [4] analyzed different miRNA biomarkers for identifying epithelial carcinomas of breast and hematological malignancies. The disease management method is used for the development of biomarkers, which could enhance early cancer detection, and it also improves the patient stratification and therapy response prediction. Some of the features such as convenience and remarkable stability increase their potential as disease biomarkers. The authors summarized the recent findings in the cancer related circulating miRNAs and presented circulating miRNAs present in plasma or serum are connected with diagnosis and/or prognosis of both primary and metastatic cancers. This review helps to understand the association between serum miR-16, miR-145, miR-155 levels, and breast cancer risk, and found a correlation of miR-155 to progesterone receptor status.

Abbas Toloie Eshlaghy et al [2] review a related work, on three classification models (C4.5 DT, SVM, and ANN), and explains the methodology used to conduct the prediction with experimental results. The researchers analyzed breast cancer data using three classification techniques to predict the recurrence of the cancer and then compared the results. The execution of three different data mining methodology is carried out using WEKA tool. To estimate validation of the models, accuracy, sensitivity, and specificity are used and compared. The results indicated that SVM is the best classifier predictor with the test dataset. Some important variables such as S-phase fraction and DNA index were not included because of their unavailability which may have decreased the performance of the models. Further studies should be conducted to improve the performance of these classification techniques by using more variables and choosing for a longer follow-up duration.

Juhyeon Kim et al [3] analyzed three cancer predictive foci for Breast cancer. The major clinical problem associated with the breast cancer is predicting its outcome (the survival or death) after the onset of therapeutically resistant disseminated illness. The clinical evident metastasis has already occurred by the time the primary tumor is diagnosed. In general, the treatments such as chemotherapy, hormone therapy, or a combination are considered to reduce the spread of breast cancer because they decrease distant metastases by one-third. Three predictive foci are related to cancer prognosis: the prediction of cancer inclination (risk assessment); the prediction of cancer recurrence (redevelopment of cancer after resolution) and the prediction of the cancer survivability. In the third case, research is focused on the predicting of the outcome in terms of life expectancy, survivability, progression, or tumor-drug sensitivity after the diagnosis of the disease. In this paper, the method focused on survivability prediction, which involves the use of methods and techniques for predicting the survival of a exacting patient based on historical data [10].

Turgay Ayer et al [1] described the purpose of using ANN algorithm to train the large prospectively collected dataset. Consecutive mammography findings can discriminate between benign and malignant disease and accurately predict the probability of breast cancer for individual patients. The collected mammography findings matched with the Wisconsin State Cancer Reporting System. The authors trained and tested their ANN by using 10-fold cross-validation to predict the risk of breast cancer and used area under curve sensitivity, and specificity to evaluate discriminative performance of the radiologists and their ANN. They assessed the accuracy of risk prediction (i.e., calibration) of their ANN by using the Hosmer-Lemeshow (H-L) goodness-of-fit test. From a clinical standpoint, ANN may be valuable because it provides an accurate post-test probability for malignancy. ANN using standardized demographic risk factors and mammographic findings, performed well in terms of both discrimination and calibration. They aim to determine whether the performance is primarily attributable to the characteristics of the training attributes of the model but ROC curve analysis is valuable to evaluate the discriminative ability of CADx models. The promising

results indicate that ANNs have the potential to help radiologists improve mammography interpretation.

From the above analysis, it is found that various methods like Decision Tree, Artificial Neural Network and Support Vector Machine (SVM) predict the cancer results accurately. However, it provides poor prognosis results in few cases. It is proposed to use SVM and Fuzzy Rule Based Classification to survival prediction.

3. EXISTING SCENARIO

Support Vector Machines (SVM) is a supervised learning model used to analyze the data and recognize patterns and also used for classification and regression analysis with associated learning algorithms. The basic SVM gets the set of input data and predicts for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, SVM training algorithm builds a model that assigns new examples into one category or the other. A SVM model is a demonstration of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. More formally, a SVM constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks.

Hyper plane separates the data into two classes with the maximum-margin. A maximum-margin hyper plane divides the training examples, such that the distance among the hyper plane and the closest examples (the margin) is maximized. Hence the selected features are classified based on the training models which produce the prediction rate of cancer survival rate.

Algorithm for SVM

Step 1: Initialize vector and class to train in SVM process.

Step 2: Classify the vector and bias using f (x_i), If y_i f (x_i) less than 1 to find the features and add a known data x_i.

Step 3: Calculate the actual size and predict the values. If the prediction is wrong then process is retrained.

Step 4: Repeat the process.

The breast cancer dataset X=(x1, y1),……,(xn, yn), C

//x and y –labeled samples and C-categories

1 Initialize vector v=0, b=0; class
// v-vector and b-bias
2 Train an initial SVM
For each $x_i \in X$ do
// xi is a vector containing features describing example i
Classify x_i using f (x_i)
If y_i f (x_i) < 1
// prediction class label
Find w', b' for known features
// w', b' for new features
Add x_i to known data
If the prediction is wrong then retrain
classlist = unique(actual);
If size(actual,1)~=size(predict,1)
predict=predict';
End
End Process

The preprocessing is initially performed in SVM process and then classification process is performed. The breast cancer dataset is collected from computational cancer biology data. The breast cancer patient's data are collected from clinical information data is stored in the database. Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome. The strongest predictors for metastases (for example, lymph node status and histological grade) fail to classify accurately breast tumors according to their clinical behavior. The fuzzy classification approach is applied to analysis the clinical cancer dataset more accurately. Breast cancer dataset http://ccb.nki.nl/data/ which contains the number of attributes like Posnodes, EVENTmeta, EVENTdeath, TIMEsurvival, TIMErecurrence, TIMEmeta, ESR1 and NIH is used. The SVM algorithm is used project to classify the survival breast cancer rate based on the frequent occurrences such as x and y terms, where xi is a vector containing features describing example i, and each yi is the class label for that example. Repeat the process for all features in the cancer dataset.
Potential drawbacks of the SVM include:

- Uncalibrated class membership probabilities.
- It is only directly applicable for two-class tasks. Therefore, to reduce the multi-class task, several binary classification problems have to be applied.
- Parameters of a solved model are difficult to interpret.

4. PROPOSED SCENARIO

Fuzzy Rule Based System (FRBS) is also named as fuzzy quantification subsethood based algorithm for classification purposes. Fuzzy rule-based modeling is a multi-model approach in which individual rule acts like a local model, fuzzy rules are combined to describe the

behavior of the system. In the proposed system, the fuzzy rule based system is used to classify the heterogeneous features and predict the survival rate in the cancer dataset more effectively.

Fuzzy classification is the process of grouping elements into a set; the membership function is defined by the truth value of a fuzzy propositional function. Fuzzy Rule Based System **(**FRBS) is also known as fuzzy inference systems or simply fuzzy systems. They are based on the fuzzy set theory, which aims at representing the set of fuzzy IF-THEN rules. Instead of using crisp sets as in classical rules, as in fuzzy rules use fuzzy sets. Rules were initially derived from human experts through knowledge engineering processes. Fuzzy sets allow for degrees of set membership, defined by values between zeros and ones. A degree of ones mean that an object is a member of the set, a value of zeros mean it is not a member, and in-between a value shows a partial degree of membership.

4.1 Mamdani Model

The grade membership of a given element is defined by the so-called membership function. The Mamdani model is built by linguistic variables in both the ancestor and following parts of the rules. So, that considering the multi-input and single-output (MISO) systems,

The Fuzzy IF-THEN rules are in the following form:

IF X1s A1 and . . . and Xn is An THEN Y is B, **(1)**

where Xi and Y are input and output linguistic variables respectively and Ai and B are linguistic values.

FRBS means defining all of its components, especially the database and rule base of the knowledge base. The operator set for the inference engine is selected based on the application or kind of model. For example, minimum or product is common choices for the conjunction operator. But the part that requires the highest effort is the knowledge base.

There are two different strategies to build FRBSs, depending on the information. The first strategy is to get information from human experts. It means that the knowledge of the FRBS is defined manually by knowledge engineers, who interview human experts to extract and represent their knowledge. However, there are many cases in which this approach is not feasible. The second strategy in FRBSs is to obtain rules by extracting knowledge from data by using learning methods.

For the components of the FRBSs that need to be learned or optimized, the following has to be performed:

Rule base: Qualified antecedent and consequent parts of the rules need to be obtained, the number of rules needs to be determined and the rules have to be optimized.

Database: Optimized parameters of the membership functions have to be defined.

Weight of rules: Especially for fuzzy rule-based classification systems, optimized weights of each rule have to be calculated.

4.2 Fuzzy subsethood measures

Let A and B be two fuzzy sets defined as the universe U. The fuzzy subsethood value of A with regard to B, S (B, A) represents the degree to which A is a subset of B:

$$\mathbf{S\ (B,\ A)} = \frac{\sum_{X \epsilon U} \nabla(\mu_B(X), \mu_A(X))}{\sum_{X \epsilon U} \mu_B(X)} \qquad \mathbf{(2)}$$

Where S (B, A)$\in$ [0, 1] and ∇ is t-norm.

The above definition of fuzzy subsethood values can be extended to calculate the degree of subsethood for linguistic terms in an attribute value V to a decision class D. If $\{A_1, A_2 \ldots . A_n\} \in$ V, it is possible to replace A with Ai and B with D in equation (2).

The generation of fuzzy rules is depending on the fuzzy subsethood values between the decision to be made and the terms of the conditional attributes. Any linguistic term that has a subsethood value that is greater than or equal to will automatically be chosen as an antecedent for the resulting fuzzy rules. The methodology, termed as the subsethood-based algorithm (SBA), assume that all piece of information gathered from the training data are equally important.

For this reason, a weighted subsethood-based algorithm (WSBA) has been proposed, in which the certain weighting strategy have been taken to represent the degree of importance. In particularly, the weights are created on the subsethood values to provide multiplication factors for each philological variable. They are calculated in an intermediate steps (steps between b and c, mentioned above) using the following formula:

$$\mathbf{W\ (D,\ } A_i \mathbf{)} = \frac{S(D, A_i)}{max_{j=1 \ldots l} S(D, A_j)}, \quad \mathbf{i=1,\ldots..l} \qquad \mathbf{(3)}$$

Where $A_i \in \{A_{1,\ldots..} A_l\}$ is the ith linguistic term of the linguistic variable A and D is the classification terms. The advantage of this method compared to the previous one is that it does not require any threshold value. The crisp weights for each linguistic term can be considered as quantifiers.

4.3 Fuzzy quantifiers

In fuzzy quantifiers, logic can be expressed as Q(x) A(x) where Q(x) is a quantifier and A(x) is a predicate in variable x. In classical logic, the quantifier and the predicate can be represented by crisp sets. In fuzzy logic, the quantifier may be applied on crisp or fuzzy sets. A quantifier based on fuzzy sets seems to be more suitable for quantifier based fuzzy models which are described in natural language.

Thus in evaluating a fuzzy quantified proposition, a quantification mechanism is needed to map the membership value $M_Q(q)$ such that:

F: $(M_Q(q))$ **-> I∈ [0, 1]**

The fuzzy quantification involves the definition of the existential quantifier∃, and the universal quantifier∀. Specifically, the quantifier is defined as

$$Q(A_{ij}, D_k) = (1-\lambda_Q)\cdot T_{\forall, A/D} + \lambda_Q \cdot T_{\exists, A/D} \quad (4)$$

Where Q is the quantifier for fuzzy set A relative to fuzzy set D and λ_Q is the degree of neighborhood of the two extreme quantifiers.

4.4 Fuzzy quantification subsethood-based algorithm

The fuzzy quantifiers are created using information extracted from data and behave as modifiers for each of the fuzzy terms. They can be used to replace the crisp weights in weighted subsethood-based algorithm. The benefits of this algorithm are

1. The use of the degree of neighborhood enables the implementation of continuous quantifiers. Thus, any possible quantifier can be created in principle.
2. The relative quantifier based method can be adapted into WSBA and it is preserved.
3. Relative subsethood values can be used as the degree of neighborhood of the fuzzy quantifiers. Thus, the two seemingly separate approaches are unified.
4. It satisfies the prosperities f quantification
5. From a clinical point of view, quantifiers are useful because their interpretability is regarded as highly important when developing decision support systems.

5. RESULTS AND DISCUSSION

In this section, the analysis has been done for both existing and proposed algorithms. It is found that the existing algorithm SVM classification shows lower performance whereas the proposed fuzzy rule based classification algorithm shows higher performance in the prediction of survival breast cancer data. The classification is analyzed using the performance metrics such as accuracy, precision and recall. From the experimental result, it is found that the proposed method provides higher performance results in terms of greater accuracy, precision, recall and reduction in time.

5.1 FUZZY RULE GENERATION

Fuzzy rule-based modeling is a multi-model approach in which individual rule acts like a local model, all rules are combined to describe the behavior of the system. In the proposed system, the fuzzy rule based system is used to classify the heterogeneous features and predict the survival rate in the cancer dataset more effectively.The fuzzy rules are defined with multiple conditional attributes and a single conclusion attribute.

The 'OR' and 'AND' are fuzzy logical operators and are interpreted by minimum and maximum operator respectively. All linguistic terms of each attribute are used to describe the ancestor of each rule initially. It may contain important information that should be taken into account. Some of such terms may be omitted due to no evaluated contribution (or with a relative weight of 0) with regard to the training data. Multiplying each linguistic term by its respective weight, the fuzzy rules to be generated will be of the form:

Rule 1 *IF* (Posnodes is low) and (EVENTmeta is high) and (TIMEmeta is low)……
THEN the class is E1
Rule 2 *IF* (Posnodes is high) and (EVENTmeta is med) and (TIMEmeta is med)……
THEN the class is E2
.
.
Rule n IF (Posnodes is high) and (EVENTmeta is high) and (TIMEmeta is high)……
THEN the class is En
In this work, the following fuzzy rules are generated.

5.2 CONFUSION MATRIX

A confusion matrix is a table that is used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

MEASURES	SVM CLASSIFIER		FUZZY RULE BASED CLASSIFIER	
	High risk 0	Low risk 1	High risk 0	Low risk 1
True Positive	59.00	22.00	65.00	25.00
False Positive	3.00	13.00	0.00	7.00
False Negative	13.00	3.00	7.00	0.00
True Negative	22.00	59.00	25.00	65.00
Precision	0.95	0.63	1.00	0.78
Recall or Sensitivity	0.82	0.88	0.90	1.00
Specificity	0.88	0.82	1.00	0.90
Accuracy	**0.84**		**0.93**	

Table 5.1 Values Extracted from Confusion Matrix

Table 5.1 depicts the value from resultant confusion matrix for SVM classifier and Fuzzy Rule based Classifier in classifying patients with low and high risk. The sensitivity and specificity values are calculated from the True Positive, False Positive, True Negative and False Negative of Confusion Matrix. All these performance metrics are discussed as follows.

Precision

Precision can be used as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. The precision is calculated as follows:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

In simple terms, high precision means that an algorithm returned significantly more related results than irrelevant results. In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as the positive class) divided by the total number of elements labeled as the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).

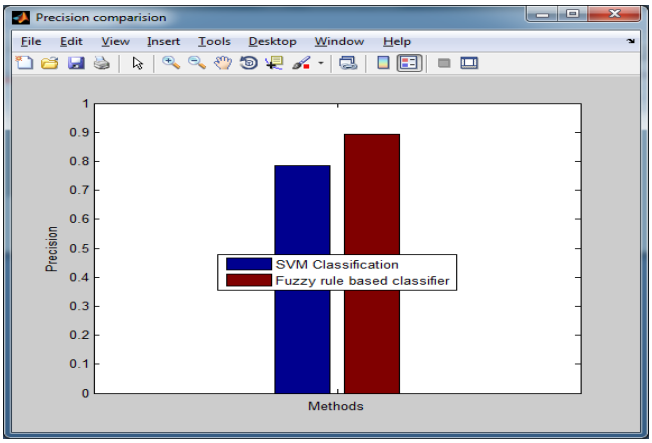


Fig 5.1. Precision Comparison

In the graph shown in Fig 5.1, x axis holds the classifiers and y axis holds the accuracy values. It is found that the precision value of existing algorithm SVM is 0.7847 which is comparatively lower than the fuzzy rule based classifier algorithm with 0.8925 precision values. From the result, it is concluded that proposed system is superior in terms of precision.

Recall

Recall is defined as the number of relevant data retrieved by a search divided by the total number of existing relevant data, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search. The calculation of the recall value is done as follows:

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

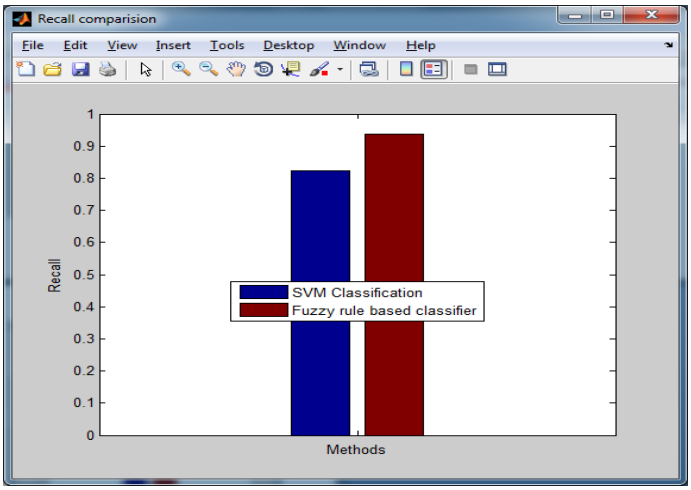


Fig 5.2. Recall Comparison

Recall is calculated as the ratio of the number of true positives divided to the total number of elements that belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belong to the positive class but should have been).

It is observed from Figure 5.2, the SVM classification algorithm gains the recall value of 0.8236 and the proposed system, fuzzy rule based classifier algorithm gains the recall value of 0.9383 which is higher than existing.

F-measure

It is a measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score.

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

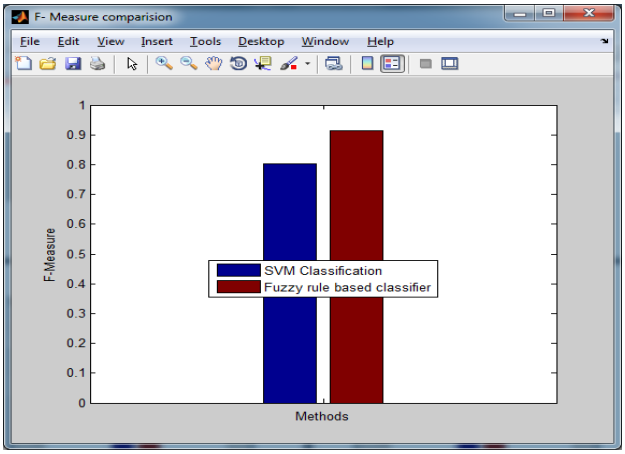


Fig 5.3. F-Measure Comparison

From Figure 5.3, it is found that the f-measure values yielded by SVM classification algorithm are 0.8037 and by fuzzy rule based classifier algorithm is 0.9149. From the result, it concludes that proposed system is superior in terms of F-measure.

Accuracy

Accuracy is a probability that the algorithms can correctly predict survival and death. Accuracy can be calculated from formula given as follows

$$Accuracy = \frac{True\ positive + True\ negative}{True\ positive + True\ negative + False\ positive + False\ negative}$$

An accuracy of 100% means that the measured values are exactly the same as the given values.

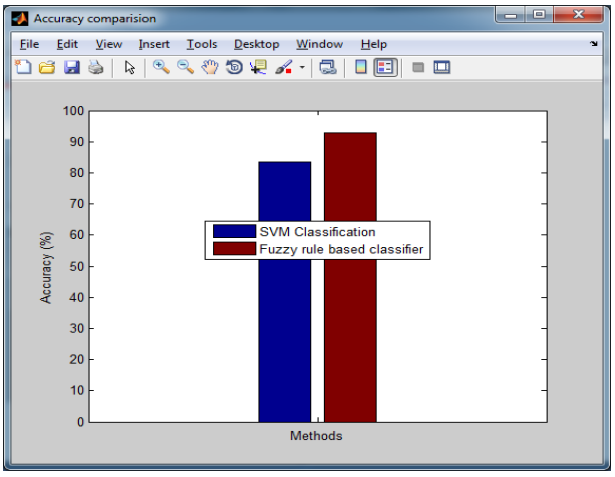
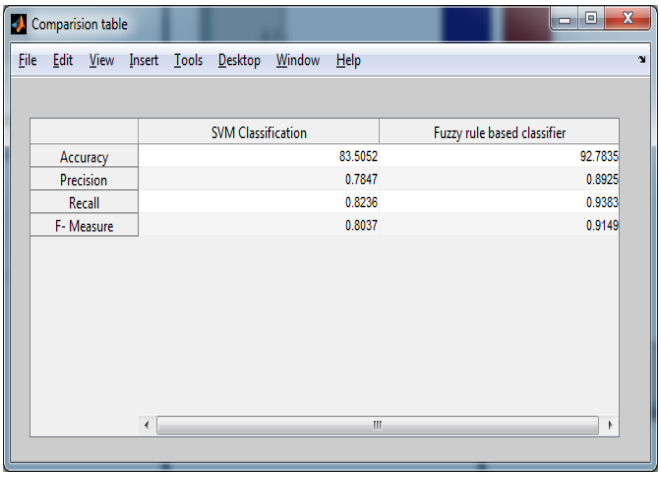


Fig 5.4. Accuracy Comparison

Figure 5.4 compares the existing and proposed system in terms of accuracy metric. The accuracy value of SVM classification algorithm with 83.5052 % accuracy is lower than the accuracy value of proposed scenario with 92.7835% accuracy. From these result, it concludes that proposed system is superior in terms of accuracy.

Comparison table



	SVM Classification	Fuzzy rule based classifier
Accuracy	83.5052	92.7835
Precision	0.7847	0.8925
Recall	0.8236	0.9383
F- Measure	0.8037	0.9149

Fig 5.5. Comparison Table

Figure 5.5 depicts the comparison of existing and proposed system in terms of precision, recall, f-measure and accuracy metric. In existing scenario, the precision, recall, f-measure and accuracy values are lower using SVM classification algorithm. In proposed system, the precision, recall, f-measure and accuracy values are higher by using the fuzzy rule based classifier algorithm. The proposed work proves its performance better than existing with 92.78% accuracy rate for breast cancer dataset classification.

FUTURE WORK

The research work can be enhanced with Optimization algorithms like Genetic algorithm and Particle swarm algorithm for efficient feature selection. Hybrid feature selection based classification approaches can be applied to increase the accuracy of cancer prediction for effective survival analysis.

CONCLUSION

The proposed work, fuzzy rule based classification algorithm aimed at applying fuzzy quantifiers in subsethood based algorithm to develop both prediction accuracy and interpretability of derived rule sets. It extracts the relevant features from the dataset and hence the size of original dataset is reduced significantly. The feature selection is performed by using fuzzy rules which selects only important features for further prediction process. Thus it takes less time for execution of implementation as well as it produces more accurate prediction results. Also, it is used to improve the overall survival rate of the diseased persons by effective prognosis. The experimental results indicates higher accuracy, precision, recall and f-measure values for the clinical breast cancer dataset and provide efficient method for handling the multidimensional heterogeneous data integration by the fuzzy rule based classification algorithm is better than the support vector machine (SVM).

REFERENCES

1. **Ayer, Turgay, et al.** "Breast cancer risk estimation with artificial neural networks revisited." *Cancer* 116.14 (2010): 3310-3321.
2. **Eshlaghy, Abbas Toloie, et al.** "Using three machine learning techniques for predicting breast cancer recurrence." *J Health Med Inform* 4.2 (2013): 124.
3. **Kim, Juhyeon, and Hyunjung Shin.** "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data." *Journal of the American Medical Informatics Association* 20.4 (2013): 613-618.
4. **Madhavan, Dharanija, et al.,** "Cancer diagnosis and prognosis decoded by blood-based circulating microRNA signatures." *Frontiers in genetics* 4 (2013).
5. **Mei-Yin C. Polley et al.,** "Statistical and practical considerations for clinical evaluation of predictive biomarkers." *Journal of the National Cancer Institute*105.22 (2013): 1677-1683.
6. **Taylor JMG, Ankerst DP, Andridge RR.** Validation of biomarker-based risk prediction models. *Clin Cancer Res*. 2011;14(19):5977–5983.
7. **URL: http://ccb.nki.nl/data/**
8. **Chang, Siow-Wee, et al**. "Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods." *BMC bioinformatics* 14.1 (2013): 170.
9. **Chuang, Li-Yeh, et al.** "Support vector machine-based prediction for oral cancer using four snps in DNA repair genes." *Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong-Kong*. 2011.
10. **Exarchos, Konstantinos P., Yorgos Goletsis, and Dimitrios I. Fotiadis.** "A multiscale and multiparametric approach for modeling the progression of oral cancer." *BMC medical informatics and decision making* 12.1 (2012): 136.
11. **Garibaldi JM, Soria D, Rasmani KA,** Consensus clustering and fuzzy classification for breast cancer prognosis. In: Proceedings 24th European conference on modelling and simulation. 2010. p. 1–4.
12. **Rakha E, Soria D, Lemetre C, Green AR, Powe DG, Nolan CC, et al.** Nottingham prognostic index plus (NPI+): a modern clinical decision making tool in breast cancer. British Journal of Cancer 2013.
13. **Cruz, Joseph A., and David S. Wishart.** "Applications of machine learning in cancer prediction and prognosis." *Cancer informatics* 2 (2006): 59.
14. **S.Kharya , D. Dubey, and S. Soni,** Predictive Machine Learning Techniques for Breast Cancer Detection, In International Journal of Computer Science and Information Technologies, Vol. 4 (6) , 2013, 1023-1028.