ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(16), DOI: 10.17485/ijst/2016/v9i16/87951, April 2016* ISSN (Online) : 0974-5645

Fuzzy based Feature Selection Scheme through Transductive SVM Technique for Cancer Pattern Classification and Prediction

J. Suganthi[1*] and V. Malathi[2]

[1]Department of CSE, Raja College of Engineering and Technology, Madurai - 625020, Tamil Nadu, India; gksuganthi123@gmail.com
[2]Anna University Regional Centre, Madurai - 625020, Tamil Nadu, India; malathi_k2000@yahoo.com

Abstract

Objectives: A reliable and precise classification of tumor types is of great importance and essential for successful Diagnosis and Drug Discovery. **Methods:** Gene expression profiling has shown a great prospective in the outcome prediction of different types of Cancers, with the innovation of Microarray Technology. Microarray cancer data, which have been organized as samples versus genes fashion, are being exploited for the tissue sample classification. They are also useful for identifying potential gene markers in each subtype of cancer, that helps to analyze a particular cancer type in a successful manner. **Findings:** In this paper a new method for classification based on Fuzzy Rough-Set Feature Selection Approach through Transductive SVM Technique, which is called as FFS+TSVM is proposed. Moreover recently, for Cancer Pattern Classification, a combined Consistency based Feature Selection Approach through Transductive Support Vector Machine (CBFS+TSVM) has proposed and its prediction accuracy has encouraged than that of existing other Classifiers. However, from the literature survey, it is revealed that the performance of the existing scheme can be improved if Fuzzy Rough Set Approach for Gene Feature Selection with Transductive Support Vector Machine is employed. The present work is implemented with the help of Bio Weka and studied thoroughly in terms of Computational Cost, Dimensionality Reduction, Threshold, Classification Error and Classification Accuracy. **Applications/Improvements:** The proposed work outperforms the existing Transductive Support Vector Machine (TSVM) in terms of Dimensionality Reduction, Threshold, Classification Error and Classification Accuracy for various Cancer Patterns.

Keywords: Fuzzy Rough Sets, Gene Selection, Information Measures, Low-Density Separation (LDS), Microarray Analysis, Semisupervised Classification, Support Vector Machines (SVM)

1. Introduction

The Cancer Classification of various tumors or cancer patterns is one of the most important Data Mining Classification Techniques for cancer pattern diagnosis and also to discover the drug. The challenge in cancer research is to achieve the highest classification accuracy during tumor classification or tumor prediction or tumor discovery. Perfect tumor type classification helps in providing good treatment, which also supports in reducing the toxicity. The innovations of microarray technologies have contributed a lot to study the gene expression more accurately under different clinical experimental setups, which provides the highest classification accuracy too[1–6]. However, to achieve the highest classification accuracy with small sample size is the challenging one[7,8]. Gene Expression based Cancer classification relies on supervised learning techniques, where the labelled data could be used for learning. The unlabelled data cannot be considered for this purpose and this data should not be considered for classification. However, a recent research has mentioned that the unlabelled data along

* *Author for correspondence*

with the labelled data produces significant improvement in classification accuracy and this type of method is called as semi-supervised learning[9,10]. Indeed, semi-supervised learning has proved to be valuable in solving different biological problems including protein classification, prediction of transcription factor–gene interaction[11], and gene-expression based on cancer subtype discovery.

The foremost research has focused on extending Support Vector Machines (SVMs) for handling semi labelled data and is based on the following idea: that is solving the standard Inductive SVM (ISVM) by treating the unknown labels as additional optimization variables. By maximizing the margin in the presence of unlabelled samples, one can trace out the decision boundary, that traverses through low density regions in the input space. In other words, this approach implements the cluster assumption for semi supervised learning with identical labels. The idea was first introduced in the name of Transductive SVM, but since it learns an inductive rule defined over the entire input space, it is also referred as Semi-Supervised SVM (S^3VM). Each cluster of samples is assumed to belong to one data class. Thus, a decision boundary is defined between clusters.

A variety of semi-supervised techniques has been proposed[12,13] and many successful algorithms directly or indirectly assume high density within class and low density between classes, but can fail when the classes are strongly overlapped[14]. This can be illustrated by comparing the well-known SVMs to their semi-supervised extensions like Transductive SVM, Progressive TSVM algorithm (PTSVM)[15], Transductive SVMs (TSVMs) and Semi-supervised SVMs (S^3VMs)[16]. Moreover, TSVMs and S^3VMs are iterative algorithms that use SVMs to gradually search a reliable hyperplane, exploiting both labelled and unlabelled samples in the training phase.

The classification of cancer, using microarray data poses another major challenge, because of huge number of features (genes) compared to the number of examples (tissue samples). This is an important problem in machine learning, which is known as Feature Selection[17].

Successful gene identification involves

- Dimension reduction to reduce computational cost.
- Reduction of noise to increase classification and performance.
- Identification of more interpretable features.

Literature survey shows that a few computational intelligence methods have been developed for gene identification. Ujjwal Maulik and et al., developed a classification system based on gene markers and this proposed technique is applied on the selected genes to classify the pattern of human cancer. To identify or design gene markers, a forward greedy reduction technique is proposed. This is a combined Consistency based Feature Selection Approach through Transductive Support Vector Machine (CBFS+TSVM). The Prediction Accuracy of this model was higher than that of existing systems. However this accuracy can be further increased by modifying the procedure of Consistency based Feature Selection approach by Fuzzy Rough Sets.

This Research work has developed an efficient Fuzzy Rough-Set based Feature Selection Approach through Transductive SVM Technique, which improved prediction accuracy.

2. Related Work

The prime objective of Semi-Supervised learning is to employ unlabelled data jointly with labelled information to improve the performance of classification. The purpose of designing Support Vector Machine (SVM) is to handle labelled data sets. It is noted that when maximizing the margin of SVM with unlabelled data, the decision boundary could be identified with respect to input space. In other words, this approach is the implementation of cluster assumption for Semi-Supervised learning which has similar labels within the data cluster[18], popularly known as Transductive Support Vector Machine (TSVM). However, as it is learning inductive rules which are defined based on the whole input space; this approach is referred as Semi-Supervised SVM (S^3VM).

2.1 S^3VM

In this section, the same model discussed by authors Ujjwal Maulik and et al., is taken and examined. The problem of binary classification is considered with the training set which consists of l labelled examples $\{(x_i, y_i)\}_{i=1}^{l}, y_i = \pm 1$, and unlabelled examples $\{x_i\}_{i=l+1}^{n}$, with n=l+u. In S^3VM classification, the minimization problem has been solved by the parameters of hyper plane (w,b) and $y_u := [y_{l+1} \ldots y_n]^T$ which is the label vector.

$$\min_{(w,b),y_u} I(w,b,y_u) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} V(y_i, o_i) + C^* \sum_{i=l+1}^{n} V(y_i, o_i) \quad (1)$$

The minimization Problem (1) is solved under the following class balancing constraint

$\frac{1}{u}\sum_{i=i+1}^{n}\max(y_i,0)=r$ or equivalently $\frac{1}{u}\sum_{i=i+1}^{n}\max(y_i=2r-1$

This class balancing constraint does help to avoid the unbalanced solutions by applying a certain user-specified fraction r. There are two broad strategies to minimize I, and those are discussed in the following sub sections.

2.2 Combinatorial Optimization

Let us consider the given fixed y_u and the optimization over (w,b) which is the training of SVM standard. Let us define

$$\Im(y_u=\min_{w,b} I(w,b,y_u)\,. \qquad (2)$$

The goal now is to minimize $\Im$ over a set of binary variable sets. There is no known algorithm that finds the global optimum efficiently.

2.3 Continuous Optimization

Let us consider for a fixed (w,b), $\text{argmin}_y V(y,o) = \text{sign}(o)$. Hence, the optimal y_u is given by the signs of $o_i = w^T x_i + b$. By eliminating y_u in this manner, a continuous objective function over (w,b) is given as

$$\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\max(0,1-y_i o_i)^2+C^*\sum_{i=l+1}^{n}\max(0,1-|o_i|)^2 \qquad (3)$$

The first two terms in the above equation are the model of standard SVM. This type of optimization problem demonstrates the implementation of the cluster assumption in S^3VMs.

2.4 Transductive SVMs for Semi Supervised Classification

The choice of Transductive samples has been done by the process of filtering the unlabelled samples as follows. Consider a binary classification problem, where the classification technique has begun with training the Support Vector Machine classifier through the available working set $W^{(0)}$. As support vectors i.e., patterns belonging to the margin $M=\{x\,||\,w^{(i)}.\phi(x)+b^{(i)}|\leq 1\}$ are the only patterns, which affect the position of the discriminant hyperplane unlabelled samples. To select these samples, N± is defined to be the positive and negative patterns within the margin bounds.

At each iteration, N± Transductive samples are selected on either side of the separating hyperplane to define the positive and negative candidate sets B±. In other words, the entire positive and negative semi labelled samples are selected from both the upper (positive) and the lower (negative) side of the margin.

A Transductive set $B_t^{(i)}=B^+\bigcup B^-$ is formed at the first (i=1) iteration. When $A_t^{(0)}=\phi$ $B_t^{(1)}$ is merged with the initial working set and the classifier is retrained and the process is repeated. Similarly the iteration is repeated for $A_t^{(1)}$ and $B_t^{(2)}$ and so on. The algorithm is illustrated below.

Input
Labelled points: S = [(x_j ,y_j)], j = 1, 2,...,l and unlabelled points: V = [(x_j)] , j = l + 1,..., n.
Output
The Transductive Support Vector Machine classifier with original set of the Transductive and the training set of original training
{
Initialize Current $W^{(0)}$ = S, $A_t^{(0)}=\phi$ (previous Transductive set) and specify C and C*
Train Support Vector Machine with $W^{(0)}$
Get label vector V (Unlabelled).
for i = 1 to T {
Select N + and N- which is the positive and negative samples of Transductive
Select positive candidate set B+ which has N+ samples of positive Transductive and similarly the set B– has N–
$B_t^{(i)}$= $B^+\cup B-$
Update the training set:
If { $A_t^{(i-1)}=\phi$
$W^{(i)}=W^{(i-1)}\cup B_t^{(i)}$
$D_t^{(i)}=B_t^{(i)}$
else
$D_t^{(i)}=A_t^{(i-1)}\cap B_t^{(i)}$
$W^{(i)}=(W^{(i-1)}-D_t^{(i-1)}\cup D_t^{(i)}$
}

8. $A_t^{(i)}=B_t^{(i)}$
9. Train Transductive Support Vector Machine with the updated $W^{(i)}$training set
10. Find the unlabelled vector set V
}
}

The above described procedure has improved classification accuracy and classifier capability as well.

2.5 Feature Selection Technique

This section, briefly describes the procedures of Consistency-Based Feature Selection (CBFS). This is used to achieve dimensionality reduction. The main objective in classification is to find the best subset so that the

classification accuracy could be increased and at the same time, data size will be reduced. The three basic steps of the Feature Selection Technique are discussed below.

- Methodology to generate the feature subject of next candidate.
- A stopping criterion to decide when to stop.
- An evaluation function to evaluate the candidate subset.

3. Identified Problem

In the previous section, the research work has discussed two popular Support Vector Machines namely Semi-Supervised Support Vector Machines S^3VM and Transductive SVMs (TSVM) for Semi Supervised Classification. From the Experimental Study, it is noted that the Transductive SVMs (TSVM) for Semi Supervised Classification is performing well for predicting and classifying Cancer Patterns. In this Transductive Support Vector Machine, according to a Transductive process, the hyperplane is defined. This defined hyperplane has integrated the unlabelled samples with the training samples, which is achieving fruitful result.

However, if both the unlabelled and labelled data follow dissimilar distributions, the integrating unlabelled data might lead to poor performance. That is the Consistency-Based Feature Selection (CBFS) approach fails to perform well for different distributions of Labelled and Unlabelled Data. This is the major issue and hence this work wanted to introduce Fuzzy Rough Set Theory to find the best gene markers. In other words, this research work has introduced Fuzzy Rough Set based Transductive SVM to achieve the higher performance when compared with the existing technique.

4. Proposed Fuzzy Feature Selection based Transductive SVM

In microarray data analysis, the data set may contain a number of redundant genes with low relevance to the classes. The presence of such redundant and non-relevant genes leads to a reduction of useful information. Preferably, the selected genes should have high relevance with the classes, while the redundancy among them should be as low as possible. The gene with high relevance is expected to predict the classes of the samples.

However, the prediction capability is reduced, if many redundant genes are selected. To assess the effectiveness of the genes, both relevance and redundancy need to be quantitatively measured. An information-measure based gene selection method (Feature Selection Approach) is presented to improve the Prediction/classification Accuracy.

4.1 Gene Selection using Information Measures

Let $G = \{G_1, ..., G_i, ..., G_j, ..., G_d\}$ denote the set of genes or fuzzy condition attributes of a given microarray dataset and S be the set of selected genes. Define $\tilde{f}(G_i, D)$ as the relevance of gene G_i (fuzzy condition attribute) with respect to class D (fuzzy decision attribute) and $\tilde{f}(G_i, G_j)$ as the redundancy between two genes G_i and G_j (fuzzy condition attributes). The total relevance of all selected genes is therefore given by

$$\Im_{relev} = \sum_{G_i \in S} \tilde{f}(G_i, D) \tag{4}$$

While the total redundancy among the selected gene is

$$\Im_{redun} = \sum_{G_i, G_j \in S} \tilde{f}(G_i, G_j) \tag{5}$$

Therefore, the problem of selecting a set S of non redundant and relevant genes from the whole set of genes G (condition attributes) is equivalent to maximizing $\Im relev$ and minimizing $\Im redun$, that is, to maximize the objective function $\Im$, where

$$\Im = \Im_{relev} - \Im_{redun} = \sum_{i} \tilde{f}(G_i, D) - \sum_{i,j} \tilde{f}(G_i, G_j) \tag{6}$$

The above discussed Gene Selection Problem can be solved using Greedy Technique, which is shown below.

1. *Initialize* $G \leftarrow \{G_1, ..., G_i, ..., G_j, ..., G_d\}$, $S \leftarrow \emptyset$.
2. *Generate FFS* M_{Gi} *for each gene* $G_i \in G$.
3. *Calculate the relevance* $\tilde{f}(G_i, D)$ *of each gene* $G_i \in G$.
4. 4. *Choose Gene* G_i *as the first highest relevance* $\tilde{f}(G_i, D)$ *In effect* $G_i \in S$, *and* $G = G \setminus G_i$.
5. *Generate resultant FFS* $M_{G_iG_j}$ *between selected gene* G_i *of S and each of the remaining genes* G_j *of G.*
6. *Calculate the redundancy* $\tilde{f}(G_i, G_j)$ *between selected genes of S and each of the remaining genes of G.*

7. *From the remaining genes of G, select gene G_j that maximizes $\tilde{f}(G_i,D)-\frac{1}{|S|}\sum_{G_i\in S}\tilde{f}(G_i,G_j)$ As a result of that, $G_j \in S$, and $G = G \setminus G_j$*
8. *Repeat the aforementioned three steps until the desired number of genes are selected.*

The relevance of $\tilde{f}(G_i,D)$ and the redundancy $\tilde{f}(G_i,G_j)$ is calculated using V -information and χ2-information which are known as information measures methods.

4.2 Generation of Fuzzy Equivalence Classes

In the proposed gene selection method, the π function in the 1-D form is used to assign membership values to different fuzzy equivalence classes for the input genes. A fuzzy set with membership function $\pi(x,c,\sigma)$ represents a set of points clustered around $\bar{c}$ and c × n FFS M_{G_i} corresponding to the i[th] gene G_i, can be calculated from the c fuzzy equivalence classes of the objects x = {x_1,..., x_j,...,x_n}, where

$$m_{kj}^{G_i} = \frac{\pi(x_j;\bar{c}_k,\sigma_k)}{\sum_{l=1}^{c}\pi(x_j;\bar{c}_l,\sigma_l)} \quad (7)$$

Corresponding to three fuzzy sets, i.e., low, medium, and high (c = 3), the following relations hold:

$$\bar{c}_1 = \bar{c}_{1ow}(G_i)\ \ \bar{c}_1 = \bar{c}_{medium}(G_i)\ \ \bar{c}_1 = \bar{c}_{high}(G_i) \quad (8)$$

$$\sigma_l = \sigma_{low}(G_i)\ \ \sigma_2 = \sigma_{medium}(G_i)\ \sigma_3 = \sigma_{high}(G_i)\ (9) \quad (9)$$

T***he F****uzzy based Feature Selecti*$_o$*n A*$_l$*gorithm is given below*

Input
Labelled points: S = [(xj ,yj)] , j = 1, 2,...,l and unlabelled points: V = [(xj)] , j = l + 1,..., n.
Output
Fuzzy Feature Selection based Transductive Support Vector Machine classifier
Begin

1. Calculating fuzzy measures
2. *Initialize various sets $W^{(0)} = S$, $A_t^{(0)} = \emptyset$ and specify C* and C*
3. *Train SVM Classifier with the working set $W^{(0)}$*
4. *Get the label vector of the unlabelled set V.* ***for i = 1 to T*** *// T is the number of iterations*
5. *Select N + positive transductive samples from the upper side of the margin and* N− *negative transductive samples from the lower side respectively.*
6. *Select positive candidate set B+ containing N+ positive transductive samples and negative candidate set B− containing N− negative transductive samples respectively.*
7. $B_t^{(i)} = B^+ \cup B-$
8. *Update the training set:*
   If $A_t^{(i-1)} = \phi$

$W^{(i)} = W^{(i-1)} \cup B_t^{(i)}$
$D_t^{(i)} = B_t^{(i)}$
else
$D_t^{(i)} = A_t^{(i-1)} \cap B_t^{(i)}$
$W^{(i)} = (W^{(i-1)} - D_t^{(i-1)} \cup D_t^{(i)}$
endif
9. $A_t^{(i)} = B_t^{(i)}$
10. *Train Transductive Support Vector Machine with the updated training set $W^{(i)}$*
11. *Obtain the unlabelled vector set V*

endfor
endfor

5. Experimental Setup and Performance Evaluation

The proposed Fuzzy Feature Selection based Transductive SVM is implemented with BioWeka and it is studied thoroughly in terms of Computational Cost, Dimensionality Reduction, Threshold, Classification Error and Classification Accuracy for various Cancer Patterns. The Bio-Weka is configured with the designed VC++ Tool for integrating developed tool with BioWeka.

The proposed Fuzzy Feature Selection based TSVM is implemented successfully and compared with the recently proposed Consistency based Feature Selection Approach through Transductive Support Vector Machine (CBFS+TSVM). From the prediction accuracy, it has encouraged that the proposed classifier outperforms the existing model. The experimental results are shown in Figures 1-5 and the Performance Analysis of the Proposed Technique on Various Cancer Patterns are shown in Table 1.

The Proposed Model is implemented using Windows 7 Home Premium 64 bit with 2.5 GHz Intel Core i5-2450M, 4 GB RAM, 64KB L1 Primary Cache and 256 KB L2 Secondary Cache.

For this experimental study, this research work has used few cancer patterns namely Bladder Cancer, Breast Cancer and Colon Cancer.

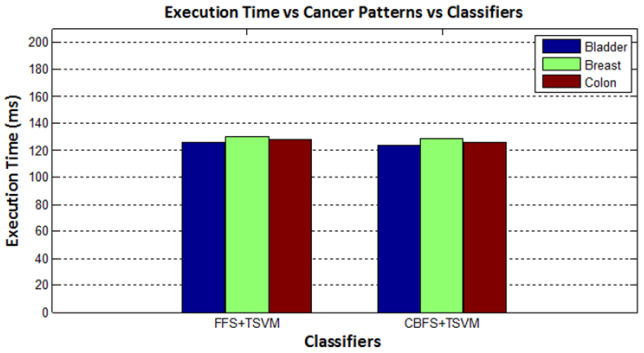

Figure 1. Execution time of proposed FFS based TSVM.

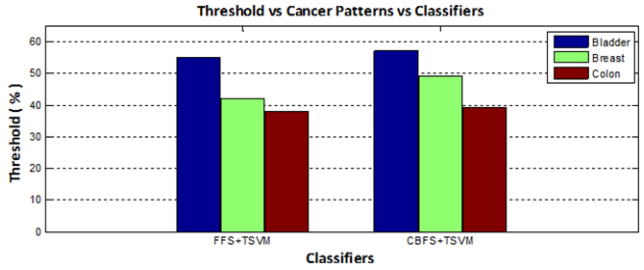

Figure 2. Threshold level to predict the cancer pattern of proposed FFS based TSVM.

This research work has compared the Execution Time of the Proposed Work FFS+TSVM with the Existing Classifier CBS+TSVM. From the Figure 1, this research work has observed that the Execution Time of both Proposed Classifier and Existing Classifier CBFS+TSVM is almost same which shows that the complexity of the proposed model is same as existing one. It has also compared the Threshold value of existing and proposed works to predict the Cancer pattern with minimum size of the sample. As shown in the Figure 2, the proposed work predicts the cancer pattern with minimum threshold when compared with the existing one. Similarly the proposed work outperforms the existing model in terms of Classification Error and Classification Accuracy, which is shown in Figure 3 and Figure 4.

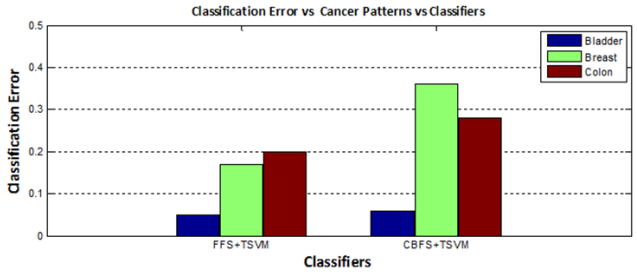

Figure 3. Classification error of proposed FFS based TSVM.

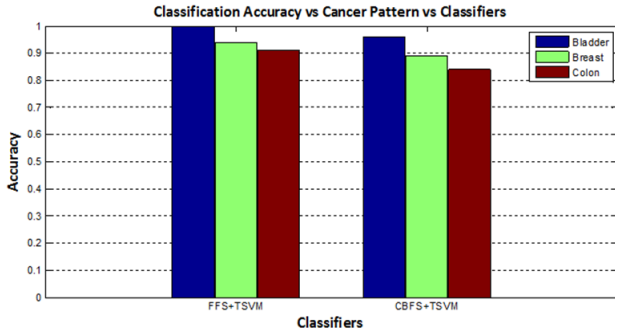

Figure 4. Classification accuracy of proposed FFS based TSVM.

Table 1 shows the Comparative Analysis of Proposed and Existing classifiers in terms of Precision, Sensitivity and Specificity for gene expression datasets, which is shown in Figure 5.

Table 1. Performance analysis of proposed technique on various cancer patterns

Accuracy Measures	Classifiers					
	Proposed FFS+TSVM			CBFS+TSVM		
	Bladder	Breast	Colon	Bladder	Breast	Colon
Precision	0.96	0.89	0.88	0.91	0.84	0.83
Sensitivity	0.97	0.94	0.91	0.92	0.89	0.87
Specificity	0.94	0.96	0.91	0.88	0.95	0.86

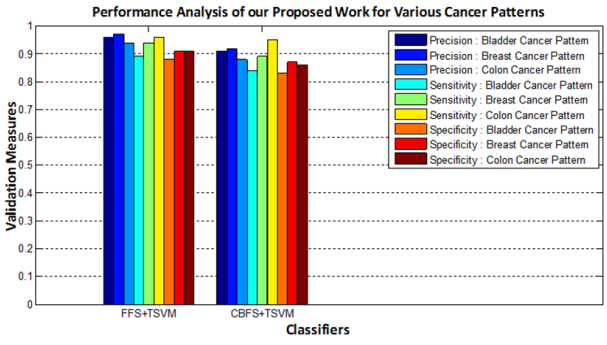

Figure 5. Validation measures of the proposed FFS based TSVM for various cancer patterns.

6. Conclusion

This Research Work has studied the recently proposed two popular classifiers namely Semi-Supervised Support Vector Machines S^3VM and Consistency Based Feature Selection approach through Transductive Support Vector Machine (CBFS+TSVM) for Semi Supervised Classification. From the experimental study, the research work has noted that the CBFS+TSVM are performing well for predicting and classifying Cancer Patterns. However, to improve the prediction accuracy further, the study has modified the Feature Selection Approach of TSVM by proposing Fuzzy Rough Set based Feature Selection Approach. Thus the research work is implemented with Bio Weka and studied thoroughly in terms of Computational Cost, Dimensionality Reduction, Threshold, Classification Error and Classification Accuracy. From the experimental results, it is observed that the proposed work outperforms the existing Transductive Support Vector Machine (TSVM) in terms of Dimensionality Reduction, Threshold, Classification Error and Classification Accuracy for various Cancer Patterns.

7. References

1. Maulik U, Mukhopadhyay A, Chakraborty D. Gene-expression-based cancer subtypes prediction through feature selection and Transductive SVM. IEEE Transactions on Biomedical Engineering. 2013; 60(4):1111–7.
2. Maji P, Pal SK. Fuzzy–rough sets for information measures and selection of relevant genes from microarray data. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics. 2010; 40(3):741–52.
3. Chandrasekhar T, Thangavel K, Sathishkumar EN. Verdict accuracy of quick reduct algorithm using clustering, classification techniques for gene expression data. International Journal of Computer Science issues. 2012; 9(1):357–63.
4. Chapelle O, Sindhwani V, Keerthi SS. Optimization techniques for semi-supervised support vector machines. Journal of Machine Learning Research. 2008; 9:203–33.
5. Bandyopadhyay S, Mitra R, Maulik U, Development of the human cancer microRNA network. BMC Silence. 2010; 1(6):245–52.
6. Mukhopadhyay A, Bandyopadhyay S, Maulik U. Multi-class clustering of cancer subtypes through SVM based ensemble of pareto-optimal solutions for genemarker identification. PLoS ONE. 2010; 5(11):1–14.
7. Kalaiselvi C, Nasira GM. Prediction of heart diseases and cancer in diabetic patients using data mining chniques. Indian Journal of Science and Technology. 2015 Jul; 8(14):1–DOI: 10.17485/ijst/2015/v8i14/72688.
8. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proceedings of the National Acad Sci USA. 2006; 103(15):5923–8.
9. Bandyopadhyay S, Maulik U, Roy D. Gene identification: Classical and computational intelligence approaches. IEEE Trans Syst, Man, Cybern C. 2008 Jan; 38(1):55–68.
10. Chapelle O, Sindhwani V, Keerthi SS. Optimization techniques for semi-supervised support vectors. J Mach Learn Research. 2008; 9(6):203–33.
11. Ernst J, Beg QK, Kay KA, Balazsi G, Oltvai ZN, Bar-Joseph Z. A semi-supervised method for predicting transcription factor–gene interactions in Escherichia coli. Plos Computational Biology. 2008; 4(1):156–61.
12. Maulik U, Bandyopadhyay S, Mukhopadhyay A. Multiobjective genetic algorithms for clustering. Applications in Data Mining and Bioinformatics. New York: Springer-Verlag; 2011.
13. Johnson R, Zhang T. On the effective Laplacian normalization for graph semi-supervised learning. J Machine Learning Research. 2007; 8:1489–517.
14. Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from examples. Journal of Machine Research. 2006; 7:2399-434.
15. Chen Y, Wang G, Dong S. Learning with progressive transductive support vector machine. Pattern Recognition Letters. 2003; 34(12):1845–55.
16. Bennett K, Demiriz A. Semi-supervised support vector machines. Proc Adv Neural Inform Process Syst. 1998; 10(5):368–74.
17. Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence. 1997; 97(1/2):245–71.
18. Bandyopadhyay S, Mukhopadhyay A, Maulik U. An improved algorithm for clustering gene expression data. Bioinformatics. 2007; 23(21):2859–65.