IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 9, September 2015.
www.ijiset.com

ISSN 2348 – 7968

An Efficient Fast Clustering and Fuzzy TSVM for Cancer Classification of Gene Expression Data

T.Chandra[1], K.Saraswathi[2]

[1] Research Scholar, Department of Computer Science,

Government Arts College (Autonomous), Coimbatore-641018

[2] Assistant Professor, Department of Computer Science,

Government Arts College (Autonomous), Coimbatore-641018

***Abstract*-----In this work professionally and effectively deal with together unrelated and redundant features, and obtain a good feature subset. We achieve this through a new feature selection framework which collected of the two joined components of unrelated feature removal and redundant feature elimination. Based on the MST method, it propose a Fast clustering based feature Selection algorithm (FAST). The past obtains features related to the target concept by eliminating unrelated ones, and the latter remove pointless features from related ones via choose legislative body from similar feature clusters, and thus produce the final subset. The unrelated feature removal is complicated once the right consequence measure is classify or selected, although the redundant feature exclusion is a bit of sophisticated. In our proposed FAST algorithm, it involves 1) the structure of the minimum spanning tree from a biased complete graph; 2) the partition of the MST into a forest with each tree representing a cluster; and 3) the selection of delegate features from the clusters. The clustering based strategy of FAST has a high possibility of producing a subset of useful and sovereign features. Inside this paper it present an different method of generate attachment standards which are call iterative FTSVM (I-FTSVM). Our method generate relationship values iteratively based on the positions of training vectors relative to the TSVM decision surface itself.**

I. INRODUCTION

Gene expression refers to the stage of creation of protein molecules defined by a gene. monitor of gene appearance is one of the most elemental move toward in measuring gene appearance is to measure the mRNA in its position of proteins, because mRNA sequence hybridize with their complementary RNA or DNA sequences though this property lacks in proteins. The DNA arrays, pioneer are novel technologies that are designed to evaluate gene expression of tens of thousands of genes in a single experiment. The ability of measure gene appearance for a very large number of genes, casing the entire genome for some small organisms, raises the problem of characterize cells in terms of gene expression, that is, using gene expression to determine the fate and functions of the cells. The most fundamental of the characterization problem is that of identifying a set of genes and its term patterns that either characterize a certain cell state or predict a certain cell state in the future. Gene selection aims to find a set of genes that best differentiate biological samples of special types. The selected genes are "biomarkers," and they form a "marker panel" for analysis. Most gene selection schemes are based on binary discrimination using rank-based schemes such as information gain, which reduces the entropy of the class variables given the selected features. One critical issue in these rank-based methods is data sparseness. For example, the estimation of the traditional information gain is an empirical estimation directly on the data. Suppose we select the 11th gene for a data set.

Major research on extending support vector machines (SVMs) to handle semi labeled data is based on the following idea: solve the standard inductive SVM (ISVM) while treating the unknown labels as additional optimization variables. By maximizing the margin in the presence of unlabeled samples, one can learn the decision boundary that traverses through low density regions while respecting labels in the input space. In other words, this approach implements the cluster assumption for semi supervised learning, that samples in a data cluster have identical labels. The idea was first introduced under the name of transductive SVM, but since it learns an inductive rule defined over the entire input space, the approach is referred to as semi supervised SVM (S3VM). Each cluster of samples is assumed to belong to one data class.

ISSN 2348 – 7968

Aimed to develop a classification system by identifying potential gene markers and subsequently applying the proposed technique on the selected genes for the classification of human cancer. A forward greedy reduction algorithm was exploited to identify the gene markers. The effectiveness of the proposed technique was compared with the LDS and ISVM on the basis of overall average accuracy.

The Transductive SVM implements the cluster assumption more directly by trying to find a hyper plane which is far away from the unlabeled points. In our opinion, the rationale for maximizing the margin is very different for the labeled and unlabeled points: For the labeled points, it implements regularization. Intuitively, the large margin property makes the classification robust with respect to perturbations of the data points. For the unlabeled points, the margin maximization implements the cluster assumption. It is not directly related to regularization (in this respect, we have a different view). Consider for instance an example where the cluster assumption does not hold: a uniform distribution of input points. Then the unlabeled points convey almost no information, and maximizing the margin on those points is useless (and can even be harmful).

TSVM might seem to be the perfect semi-supervised algorithm, since it combines the powerful regularization of SVMs with a direct implementation of the cluster assumption. However, its main drawback is that the objective function is non-convex and thus difficult to minimize. Consequently, optimization heuristics like SVM light sometimes give bad results and are often criticized. The main points of this paper are:

• The objective function of TSVM is appropriate, but different ways of optimizing it can lead to very different results. Thus, it is more accurate to criticize a given implementation of the TSVM rather than the objective function itself.

• The search for a low density decision boundary is difficult. The task of the TSVM algorithm can be eased by changing the data representation

II. PROPOSED WORK

- Inductive SVM
- Consistency-Based Feature Selection
- Transductive SVMs
- Fast clustering based feature selection
- Fuzzy transductive SVM (FTSVM)
- Performance evaluation

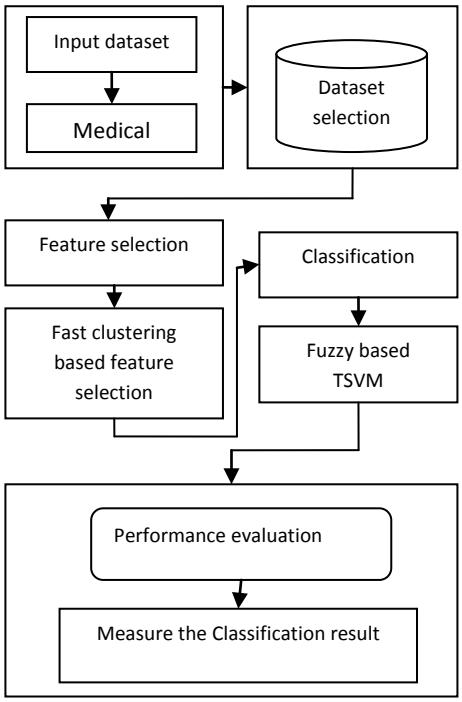


Figure 1.1 System Architecture

A. *Inductive SVM*

For a typical learning task $P(\bar{x}, y)\ P(y \mid \bar{x})P(\bar{x})$, an inductive SVM learner aims to build a decision function $fL : \bar{x} \rightarrow \{-1,+1\}$ based on a training set *train S* , which is

$$fL = L(S_{train})$$

In SVM theory, the computation of fL can be traced back to the classical structural risk minimization (SRM) approach, which determines the classification decision function by minimizing the empirical risk. For SVM, the primary concern is determining an optimal separating hyper-plane that gives a low generalization error. Usually, the classification decision function in the linearly separable problem is represented by

$$f\bar{w},b = sign(\bar{w}\cdot \bar{x} + b)$$

In SVM, this optimal separating hyper plane is determined by giving the largest margin of separation between different classes. It bisects the shortest line between the convex hulls of the two

ISSN 2348 – 7968

classes, which is required to satisfy the following constrained minimization, as

$$Minimize: \frac{1}{2}\vec{w}^T\vec{w}$$
$$Subject\ to: y_i(\vec{w}\cdot\vec{x}+b)\geq 1$$

For the linearly non-separable case, the minimization problem needs to be modified to allow misclassified data points. This modification results in a soft margin classifier that allows but penalizes errors by introducing a new set of variables as the measurement of violation of the constraints.

$$Minimize: \frac{1}{2}\vec{w}^T\vec{w}+C(\sum_{i=1}^{L}\xi_i)^k$$
$$Subject\ to: y_i(\vec{w}\cdot\varphi(\vec{x}_i)+b)\geq 1-\xi_i$$

where and k are used to weight the penalizing variables $\xi^l i=1$, and ϕ (·) is a nonlinear function which maps the input space into a higher dimensional space. Minimizing the first term corresponds to minimizing the VC-dimension of the learning machine and minimizing the second term) controls the empirical risk. Therefore, in order to solve problem, we must construct a set of functions, and implement the classical risk minimization on the set of functions. Here, a Lagrangian method is used to solve the above problem. Then, the equation can be written as

$$Minimize: F(\Lambda)=\Lambda\cdot 1-\frac{1}{2}\Lambda\cdot D\cdot\Lambda$$
$$Subject\ to: \Lambda\cdot y=0;\Lambda\leq C;\Lambda>0$$

The decision function can be rewritten as

$$f(x)=sign(\sum_{i=1}^{l}y_i\lambda_i^*(\vec{x}\cdot\varphi(\vec{x}_i)+b^*)$$

B. *Consistency-Based Feature Selection*

Feature selection is a useful technique in dealing with dimensionality reduction. In classification, it is used to find an optimal subset of relevant features so that the overall accuracy is increased while the data size is reduced. When a classification problem is defined by features, the number of features can be quite large, many of which can be irrelevant. A relevant feature can increase the performance of a classifier while an irrelevant feature can deteriorate it. Therefore, in order to select the relevant features, it is necessary to measure the goodness of selected features using a feature selection criterion. The class separability is often used as one of the basic selection criteria. In this study, consistency measure is exploited as a selection criterion that does not attempt to maximize the class separability but aims to retain the discriminatory power of the original features.

Input Decision table $R=(U,F\bigcup d,f)$.

Output One reduct red.
begin
1. $red\leftarrow\emptyset$ // *red is the pool to conserve the selected attributes.*
2. For each $a_i\in P-red$, compute $Sig(a_i,red,G)=\delta_{red}\bigcup_{a_i}(G)-\delta_{red}(G)$.
end.
3.Select the attribute a_k which satisfies: $Sig(a_k,red,G)=\max(Sig(a_i,red,E))$.
4. if $Sig(a_k,red,G)>0$
$red\rightarrow red\bigcup a_k$
goto step 2
else
return red
5.**end**

C. *Transductive SVMs*

The ISVM classifier is based on the hyper planes that maximize the separating margin between two classes using the available labeled samples. The ISVM was originally developed as a two-class pattern recognition problem . TSVMs are basically iterative algorithms that gradually search the optimal separating hyper plane in the feature space with a transductive process that incorporates unlabeled samples in the training phase.

This procedure improves the generalization capability of the classifier. Gradually, the separating hyper plane will move to a finer position in subsequent iterations. This can be explained by arguing that reducing the misclassification of transductive samples can lead to the identification of a more reliable discriminate function. However, like all semi supervised techniques, also for the proposed transductive SVM, it is not possible to guarantee an increase of accuracy with respect to the inductive SVM in all cases. The convergence of the learning depends on the similarity between the problems represented by the training points and unlabeled points. In the proposed TSVM multiclass problem, we adopted one-against-all (OAA) architecture.

Input Labeled points: S = [($\mathbf{x}_j$, y_j)], j = 1, 2, . . . , l and unlabeled points: V = [($\mathbf{x}_j$)], j = l + 1, . . . , n.

Output Transductive SVM classifier with original training set and the transductive set.

begin

1. Initialize the working set $W^{(0)}$ = S, previous transductive set $A^{(0)}{}_t$ = ∅ and specify C and C *

2. Train SVM classifier with the working set $W^{(0)}$

3. Obtain the label vector of the unlabeled set V .

for i **= 1 to** T // T **is the number of iterations**

ISSN 2348 – 7968

4. Select N+ positive transductive samples from the upper side of the margin and N− negative transductive samples from the lower side respectively.

5. Select positive candidate set B+ containing N+ positive transductive samples and negative candidate set B − containing N− negative transductive samples respectively.

6. $B^{(i)}{}_t = B^+ \cup B^-$

7. Update the training set:

if $A^{(i-1)}{}_t = \emptyset$ W(i) = $W^{(i-1)} \cup B^{(i)}{}_t$

$D^{(i)}{}_t = B^{(i)}{}_t$

else

$D^{(i)}{}_t$ = A(i−1) t ∩ B(i) t

$W^{(i)} = (W^{(i-1)} - D^{(i-1)}{}_t) \cup D^{(i)}$ t

end if

8. $A^{(i)}{}_t = B^{(i)}{}_t$

9. Train TSVM classifier with the updated training set W(i)

10. Obtain the label vector of the unlabeled set V

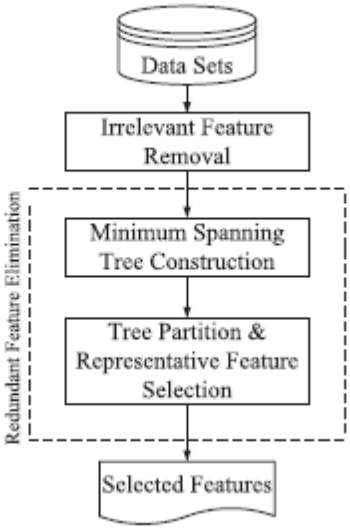
end for

end

D. Fast clustering based feature selection

They develop a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. We achieve this through a new feature selection framework which composed of the two connected components of irrelevant feature removal and redundant feature elimination. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset.

The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the MST into a forest with each tree representing a cluster; and 3) the selection of representative features from the clusters. In order to more precisely introduce the algorithm, and because our proposed feature subset selection framework involves irrelevant feature removal and redundant feature elimination



Framework of the proposed feature subset selection

The proposed FAST algorithm logically consists of three steps: 1) removing irrelevant features, 2) constructing an MST from relative ones, and 3) partitioning the MST and selecting representative features.

For a data set D with m features F = {F_1; $F_{2;}$ . . . ; F_m} and class C, we compute the T-Relevance SU {F_i; C) value for each feature $F_i \le i \le m$ in the first step. The features whose SUF_i; CÞ values are greater than a predefined threshold θ comprise the target-relevant feature subset $F' = \{F'_1, F'_2, \ldots . F'_m\}$. we first calculate the F-Correlation $SU(F'_i, F'_j)$ value for each pair of features F'_i and F'_j $(F'_i, F'_j \in F' \wedge i \neq j)$ then viewing features F'_i and F'_j as vertices and $(SU(F'_i, F'_j)\ i \neq j)$ as the weighted of the edge vertices F'_i and F'_j a weighted complete graph G=(V,E) is constructed where V=$\{F'_i | F'_j \in F' \wedge i, j \in (1, K)\}$ and $E = \{(F'_i, F'_j) | (F'_i, F'_j \in F' \wedge i, j \in (1, K) \wedge i \neq j\}$. Complete graph G reflects the correlations among all the target-relevant features. Unfortunately, graph G has k vertices and k =k /2 edges. For high-dimensional data, it is heavily dense and the edges with different weights are strongly interweaved. Moreover, the decomposition of complete graph is NP-hard. Thus for graph G, we build an MST, which connects all vertices such that the sum of the weights of the edges is the minimum, using the well known Prim algorithm.

ISSN 2348 – 7968

Input : D($F_1, F_2, \ldots\ldots\ldots F_M, C$) the given dataset θ *the* T-relevance threshold

Output: S-selected feature subset

//irrelevant feature removal

1.For i=1 to m do

2.T-relevance = SU(F_i, C)

3.If T-relevance> θ then

4.S=S∪{F_i};

//minimum spanning tress construction

5.G=NULL;G is the complete graph

6.For each pair of features {(F_i' , F_j') ⊂ S do

7. F-correlation=SU(F_i' , F_j')

8.Add F_i' *and /or* , F_j' to G with F-correlation as the weight of the corresponding edge

9. minSpanTree = Prim(G);

// tree partition and representative feature selection

10.Forest =Minspantree

11.For each edge $E_{ij} \in Forest$ do

12.if SU(F_i' , F_j') < $SU(F_i', C) \wedge SU(F_i', F_j') < SU(F_j', C)$ *then*

13. Forest =Forest -E_{ij}

14.S=∅

15.For each tree $T_i \in Forest$ do

16. F_R^j =argmax $argmax_{F_k' \in T_i}$ SU(F_k', c)

17.$S = S \cup \{F_R^j\}$;

18.Return S

E. Fuzzy transductive SVM (FTSVM)

Consider the set of training vectors x*i* belonging to a given class (+1 or −1). Let us assume that this set is sufficiently large to be a representative sample of the underlying distribution of training vectors in the form of a cloud of points in feature space. We further assume that the membership *si* of any given vector x*i* of class *di* is a function of the density of training points of that class in the vicinity of ϕ(x*i*) in feature space (ie. the lower the density of points of class *di* in the vicinity of ϕ(x*i*), the lower the membership value *si* of that point). Now, any reasonable decision surface in feature space (such as that given by the TSVM itself) should only pass through areas sparsely populated by points of either class, and moreover should correctly bisect the more densely populated regions of feature space. Hence those points lying closest to, or on the wrong side of, the decision surface (ie. the margin and error support vectors) are more likely to be outliers than points lying in more densely populated regions of feature space. Error (incorrectly classified) support vectors, in particular, may be quite reasonably expected to be outliers.

This motivates our proposed method. Let us suppose that we have constructed a non-fuzzy TSVM decision surface for our training set Θ. We then calculate memberships *si* using the equation:

$$si = h(\xi i)$$

where *h* is some continuous non-increasing function satisfying:

$$\lim_{\xi \to 0+} h(\xi) = 1$$
$$0 < h(\xi) \leq 1 \quad \forall \xi \geq 0$$

and ξi is the slack error. Using these memberships a fuzzy TSVM boundary may be constructed, and if desired the process may be iterated using this new (and presumably more accurate) decision boundary either a fixed number of times or until the membership vector **s** converges. We call our method iterative fuzzy TSVM (I-FSVM). To be specific, our method of generating membership values is as follows (noting that incremental training may be used at step 2 to minimize the re-training time for each iteration):

1) Set **s** = **1**.

2) Solve the FTSVM dual training problem.

3) For all $i \in Z_N$ set (where $0 < \mu < 1$):

$$s_i = s_i^{\text{previous}} + \mu \left[h(\xi_i) - s_i^{\text{previous}} \right]$$

4) Exit if termination condition met.

5) Otherwise repeat from step 2.

ISSN 2348 – 7968

It consider two possible termination conditions. The first (and simpler) termination condition is to stop after a fixed number (*n*) of iterations, which we will refer to as I-FTSVM *n*. The second termination condition is to continue until the rate of change of **s** becomes sufficiently small; indicating that the class membership vector **s** has converged to some value.

III. PERFORMANCE EVALUATION

The performance of the proposed method has been evaluated in terms of sensitivity, specificity and accuracy. The three indices are defined as follows.

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = TN/(TN+FP)$$

$$Accuracy = (TP+TN)/(TP+TN+FP+FN)$$

Where
True Positive (TP): the classification result is positive in the presence of the clinical abnormality.
True Negative (TN): the classification result is negative in the absence of the clinical abnormality.
False Positive (FP): the classification result is positive in the absence of the clinical abnormality.
False Negative (FN): the classification result is negative in the presence of the clinical abnormality.

IV. CONCLUSION

In this work we propose a two technique for feature selection and classification. They are fast clustering based feature selection technique and fuzzy based transductive support vector machine (FTSVM). Compared with other different types of feature subset selection algorithms, the proposed algorithm not only reduces the number of features, but also improves the performances of the well-known different types of classifiers. In contrast to standard approaches which make underlying assumptions about the distribution of training data, our method generates membership values based on their positions relative to the TSVM decision function. The proposed system has very effective rate in accuracy, specificity and sensitivity rate compared to the existing system.

REFERENCES

[1] S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik, “An improved algorithm for clustering gene expression data,” Bioinformatics, vol. 23, no. 21, pp. 2859–2865, 2007.

[2] S. Bandyopadhyay, U. Maulik, and D. Roy, “Gene identification: Classical and computational intelligence approaches,” IEEE Trans. Syst., Man, Cybern. C, vol. 38, no. 1, pp. 55–68, Jan. 2008.

[3] S. Bandyopadhyay, R. Mitra, and U.Maulik, “Development of the human cancer microRNA network,” BMC Silence, vol. 1, no. 6, 2010.

[4] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold regularization: A geometric framework for learning from examples,” Univ. Chicago, Chicago, IL, Tech. Rep. TR 2004–2006, 2004.

[5] K. P. Bennett and A. Demiriz, “Semi-supervised support vector machines,” in Proc. Adv. Neural Inform. Process Syst., 1998, vol. 10, pp. 368– 374.

[6] A. Blum and P. Langley, “Selection of relevant features and examples in machine learing,” Artif. Intell., vol. 97, no. 1/2, pp. 245–271, 1997.

[7] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” Knowl. Discov. Data Mining, vol. 2, pp. 121–167, 1998.

[8] [Online]. Available: http://www.biolab.si/supp/bi-cancer/projections/index.htm

[9] O. Chapelle, V. Sindhwani, and S. S. Keerthi, “Optimization techniques for semi-supervised support vectors,” J.Mach. Learn. Res., vol. 9, pp. 203– 233, 2008.

[10] O. Chapelle and A. Zien, “Semi-supervised classification by low-density separation,” in Proc. 10th Int. Works. Artif. Intell. Stat., 2005, pp. 57–64.

[11] Y. Chen, G. Wang, and S. Dong, “Learning with progressive transductive support vector machine,” Pattern Recognit. Lett., vol. 34, no. 12, pp. 1845– 1855, 2003.

[12] M. Dash andH. Liu, “Consistency based search in feature selection,” Artif. Intell., vol. 151, pp. 155–176, 2003.

[13] A. Dupuy and R. M. Simon, “Critical review of public microarray studies in cancer outcome and guidelines on statistical analysis and reporting,” J. Nat. Cancer 1nst., vol. 99, pp. 147–157, 2007.