A Novel Comparison of Weighting Method for Stable Gene Selection using Sampling

Dr. G. Baskar[1] Dr.P.Ponmuthuramalingam[2]
[1]Assistant Professor [2]Associate Professor & Controller of Examinations
[1,2]Department of Computer Science & Engineering
[1]KSG College of Arts & Science Coimbatore, Tamil Nadu, India [2]Government Arts College (Autonomous) Coimbatore, Tamil Nadu, India

Abstract— Data mining is a system of searching huge amounts of data for patterns. It is a comparatively new perception which is straightly related to computer science. Technically, data mining is the process of discovery associations among dozens of fields in big relational databases. Feature selection has been normally viewed as a problem of searching for an optimal subset of features guided by some evaluation measures. The identification and validation of molecular biomarkers for cancer diagnosis, prognosis, and therapeutic aims is an important problem in cancer genomics. This paper evaluates a comparison of weighting method for boosting the algorithm for finding stable gene selection.

Key words: Instance weighting, Sample Weighting, Hybrid Weighting, Gene Stability, Feature Selection

I. INTRODUCTION

Feature selection process can be decomposed into four basic steps, they are subset generation, subset evaluation, stopping criterion, and result validation. Subset generation involves a search procedure, which is used for creating candidate feature subsets for further evaluation.

Due to the time-consuming, costly, and labor-intensive nature of clinical and biological validation experiments, it is critical to select a list of high-potential biomarker candidates for validation [Pepe et al. 2001]. Gene expression microarray data [Golub et al. 1999] and relative genomic hybridization (CGH) microarrays [Pinkel et al. 1998] are commonly used for identifying candidate genes in various cancer studies. From a machine learning view, the selection of candidate genes in this context can be stared as a problem of feature selection from high dimensional labeled data. There survives various feature selection algorithms well adapted into problem of gene marker selection and they aim at improving classification accuracy while reducing dimensionality and model complexity.

This issue is mostly critical for applications where feature selection is used as a knowledge discovery tool for detecting robust and truly relevant underlying characteristic features which stay the same as training data varies.

In fact, biologists are interested in finding a small number of features (genes or proteins) that explain the mechanisms driving different behaviors of microarray samples [Pepe et al. 2001]. Biologists instinctively have high confidence in the result of an algorithm that selects alike sets of genes under some variations to the micro array samples. Even though most of these subsets are as good as each other in terms of predictive performance [Davis et al. 2006; Kalousis et al. 2007; Loscalzo et al. 2009], such instability reduces the confidence of domain experts in experimentally validating the selected features.

II. MARGIN CONCEPT

Sample Margin (SM measures how much can an instance travel before it hits the decision boundary. On the other hand, Hypothesis Margin (HM) measures how much can the hypothesis travel before it hits an instance. Margins play an important role in modern machine learning, research, and have been used together for theoretical overview bounds, as procedures for algorithm design for classifier with respect to its decisions is measure by the confidence. As described there are two natural ways of stating the margin of a sample with respect to a classifier.

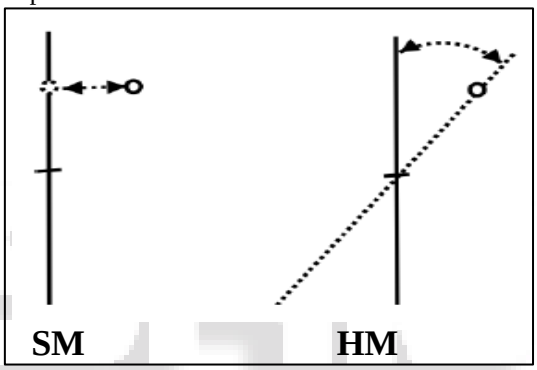


Fig. 1: Sample Margin (SM) and Hypothesis Margin (HM)

Sample margin measures the distance between a sample and the decision boundary of a classifier. Support Vector Machine (SVM), for example, uses this type of margin; it finds the separating hyperplane with the largest sample margin for support vectors. An alternative definition, hypothesis margin, measures the distance among the hypothesis of a sample and the closest hypothesis that gives an alternative label to the sample. Hypothesis margin involves a distance measure between hypotheses (classifiers). For example, AdaBoost uses this kind of margin with the L1-norm as the distance measure between hypotheses. Feature selection methods established under the large margin principles such as SVM-RFE and ReliefF assess the relevance of features according to their individual contributions to the margins. For 1-Nearest Neighbor (1NN) classifier, proved that 1) the sample margin are lower bounds by the hypothesis margin ; and 2) the hypothesis margin of a sample x with respect to a training set D can be computed by the following formula

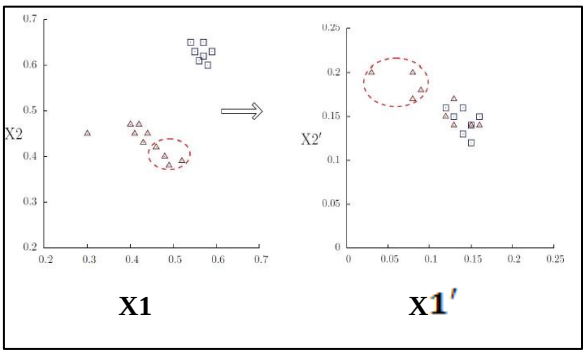
$$(\mathrm{X}) - \frac{1}{2}\left(\|\mathrm{X} - \mathrm{X}^{\mathrm{M}}\| - \|\mathrm{X} - \mathrm{X}^{\mathrm{H}}\|\right) \quad (1)$$

Where x^H and x^M represent the nearest samples (called Hit and Miss) to x in D with the similar and opposite class labels, respectively. Since hypothesis margin is easy to compute and large hypothesis margin ensures large sample margin, In this thesis the research focus on hypothesis margin.

All rights reserved by www.ijsrd.com

Let $D - \{(X_i, y_i)\}\ ^{n}_{i} = 1$ denote a training set of n labeled samples, where x_i is a sample vector in the feature space R^d state by d features $X_1; \ldots; X_d$, and the variable Y has its class value y. For gene selection, a gene expression microarray data set, consisting of the expression levels of d genes across n samples labeled by experimental conditions, can be represented as a training set for feature selection, with each gene represented by a feature.

III. MARGIN VECTOR FEATURE SPACE



A descriptive example for Margin Vector Feature Space. Each data point in the original feature space (left) is projected to the margin vector feature space (right) giving to its hypothesis margin in the original feature space. The class labels of data points are distinguished by triangles and squares.

In sample weighting for stable gene selection, the concept of hypothesis margin is implemented. By decomposing the margin of a sample along each dimension, the sample in the original feature space can be characterized by a new vector (called margin vector) in the margin vector feature space well-defined as follows

Definition 1 Let X-($X_{,\ldots\ldots,}X_d$) be a sample in the original feature space R^d and X^H and X^M represent the nearest samples to X with the similar and opposite class lables, respectively. For each X ϵ $R^{d,}$, X can be mapped to X'-(x'_j ,…., x'_d) in a new feature space R^d according to

$$x'_j - |x_j - x_j^M| - |x_j - x_j^H| \qquad (2)$$

where x'_j is the j^{th} coordinate of x' in the new feature space R^d, and x_j, x_j^M, or x_j^H is j^{th} coordinate of x, x^H, or x^M in the original feature space R^d, respectively. Vector x' is called the margin vector of X, and R^d is known as the margin vector feature space.

In essence, x' captures the local profile of feature relevance for every feature at x. The greater the value of x'_j the additional feature X_j contributes to the margin of sample x. Thus, the margin vector feature space captures local feature relevance outlines (margin vectors) for all samples in the original feature space. Figure demonstrates the idea of margin vector feature space over a 2D example. Each labeled data point (triangle or square) is a sample with two features. Each sample in the original feature space (left) is likely into the margin vector feature space (right) according to figure. It is clearly seen that samples considered with triangles exhibit mostly different outlying degrees in the two feature spaces. Exactly, those in the dashed ovals are regularly distributed within the proximity to the rest of the triangles (except the outlier on the leftmost) in the original feature space, but are obviously separated from the common of the samples in the margin vector feature space.

The outlier triangle in the original space becomes part of the majority group in the margin vector feature space. To choose the complete relevance of features X1 versus X2, one instinctive idea is to take the average over each and every margin vectors, as accepted by the well-known ReliefF algorithm. However, since the samples in the dashed oval exhibit largely separate margin vectors from the remaining of the samples, the presence or absence of these samples in the training set will disturb the global decision on which feature is more related. From the illustrative example, it can understood that the margin vector feature space captures the distance between samples with respect to their margin vectors (instead of feature values in the original space), and permits the detection of samples that mostly deviate from others in this esteem. By finding and reducing the importance on these outlying samples, more stable results can be formed from a feature selection method.

IV. MARGIN-BASE WEIGHTING ALGORITHM

The previous definition of margin vector feature space only considers one nearest neighbor from all class. To decrease the effect of noise or outliers in the training set on the transformed feature space, various nearest neighbors from all class can be used to calculate the margin vector of a sample. The research work considers all neighbors after each class for a given sample. Equation (4) can then be extended to

$$x'_j - \sum_{i=1}^{m} |x_j - x_j^{Mi}| - \sum_{i=1}^{h} |x_j - x_j^{Hi}| \qquad (3)$$

where x_j^{Hi} or x_j^{Mi} j represents the jth component of the l^{th} neighbor to x with the same or opposite class label, respectively. m or h denotes the total number of Misses or Hits (m + h equals the total number of samples in the training set excluding x).

Once the margin vector feature space is created, the next task is to exploit the inconsistency of margin vectors in this space to weight samples in the original space. To quantitatively assess the outlying degree of all margin vectorx′, and measure the average distance of x' for all other margin vectors; larger average distance specifies higher outlying degree. As illustrated in Figure 4.5, the global decision of feature relevance is more complex to samples that largely deviate from the rest of the samples in the margin vector feature space than to samples that require low outlying degrees.

To increase the stability of a feature selection method under training data variations, and assign lower weights to samples with greater outlying degrees. This choice is consistent with the intuition behind importance sampling introduced earlier. Exactly, the weight for a sample x in the original feature space is given by

$$W(X) - \frac{1\sqrt{\text{dist}(X')}}{\sum_{i=1}^{n} 1\sqrt{\text{dist}(X'_i)}}$$

Where

$$\overline{\text{dist}}(X') - \frac{1}{n-1}\sum_{i=1, X'_j \neq X'}^{n-1} \text{dist}(X', X'_i) \qquad (4)$$

All rights reserved by *www.ijsrd.com*

Margin Based Instance Weighting

Input: data $D - \{x_i\}_{i=1}^n$
Output: Weight vector W for all instances in D
// Margin Vector Feature Space Transformation
for $i - 1$ to n do
for $j - 1$ to n do
For x_i, compute $x'_{i,j}$ according to equation (3)
end for
end for
// Margin Based Instance Weighting
Calculate and store pair-wise distances among all margin vectors x'_i
for $i - 1$ to n do
For x_i, compute its weight w(x_i) (4)
end for

Fig. 1: Algorithm outlines the key steps of Margin Based Instance Weighting

For the margin-based sample weighting algorithm. Both feature space transformation and sample weighting include distance computation along every features for all pairs of samples: the previous is the original feature space, and the latter is the margin vector feature space. Then these calculations dominate the time complexity of the algorithm, the complete time complexity of the algorithm is o(n^2 * d), wherever n is the sample size and d is the number of features (genes). Therefore, the algorithm is very efficient for microarray data with small sample size.

Margin Based Sample Weighting

Input: data $D - \{x_i\}_{i=1}^n$
Output: Weight vector W for all samples in D
Feature space transformation
for $i - 1$ to n do
for $j - 1$ to n do
For x_i, compute $x'_{i,j}$ according to equation (3)
end for
end for
Sample Weighting
Calculate and store pair-wise distances among all margin vectors x'_i
for $i - 1$ to n do
For x_i, compute its weight according to equation (4)
end for

For hybrid Weighting each feature is assigned by a criterion termed Margin Fraction (MF), that contributes to the margin combined in the final output. To perform feature selection, consider features as the weak learners for boosting. Trees are among the most popular base procedures in machine learning, the predictor variables invariant under monotone transformations have the advantage, i.e. no need to search for improved data transformations.

Let F be the total number of unique features used across all T rounds, i.e. decision stumps, and for any chosen feature υ, let h(υ,j) be the decision stump corresponds to the j-th use of feature υ, and let α(υ,j) be the associated voting confidence. Let Nυ be the total number of times that feature υ is used.

$$\sum_{v=1}^{F} \boldsymbol{N_v} = \boldsymbol{T} \quad (5)$$

$$\boldsymbol{H(X_i)} \sum_{v=1}^{F} \sum_{j=1}^{N_v} \boldsymbol{\alpha(v,j)h(v,j)(x_i)} \quad (6)$$

Now for any individual feature υ, one can consider the weighted linear combination associated with that feature and the "conditional" margin associated with just that weighted linear combination for any instance i.

Margin Based Hybrid Weighting

Input: data $D - \{x_i\}_{i=1}^n$
Output: Weight vector W={w_1..... w_n}for all samples in D
Feature space transformation
for $i - 1$ to n do
for $j - 1$ to n do
For x_i, compute $x'_{i,j}$ according to equation
end for
end for
Hybrid Weighting
Calculate and store pair-wise Euclidean distances among all margin vectors x'_i
for $i - 1$ to n do
For x_i, MF_v compute its weight according to equation.
end for

V. Stability Measures

Here with the similarity-based approach where the stability of a feature selection method is measured by the average over all pairwise similarity comparisons among all feature subsets (gene signatures) obtained by the same method from different subsampling of a data set $\boldsymbol{D - \{x_i\}_{i=1}^n}$ be a set of subsampling of a data set of the same size, and r_i be the feature subset selected by a feature selection method F on the subsampling D_i The stability of F over D is

$$\overline{\boldsymbol{S_{D,F}}} \quad - \frac{2\sum_{i=1}^{q-1}\sum_{j=i+1}^{q} \boldsymbol{S(r_i,r_j)}}{\boldsymbol{q(q-1)}} \quad (7)$$

where $\boldsymbol{S(r_i,r_j)}$ represents a similarity measure between subsets $\boldsymbol{r_i,r_j}$ The stability of a feature selection method depends on the specific choice of the similarity measure $\boldsymbol{S(r_i,r_j)}$ Simple measures such as the percentage of overlap or Jaccard index can be applied.

VI. Summary of the Data Sets

The algorithm experimented with four commonly studied public gene expression microarray data sets are summarized in Table1, The Colon cancer data set has been normally used in earlier studies in gene selection and classification. It consists of the gene expression profiles of 2,000 genes for

All rights reserved by www.ijsrd.com

62 tissue samples among which 40 are colon cancer tissues and 22 are normal tissues.

The Leukemia data set is another widely used benchmark data set. It consists of gene expression profiles of two classes of leukemia: acute lymphoblastic leukemia (ALL) and acute myeloblastic leukemia (AML). The data set consists of 7,129 genes and 72 samples (47 ALL and 25 AML). The Prostate data set involves of gene expression profiles of 6,034 genes for 52 prostate tumor samples and 50 normal samples. The Lung cancer data set involves of gene expression profiles of 12,533 genes for 181 lung tissue samples among which 31 are of malignant pleural mesothelioma (MPM) and 150 are of adenocarcinoma (ADCA).

The link is http://cs.binghamton.edu/~lyu/KDD08/data/

S.no	Name	Features	Samples
1	Colon	2000	62
2	Leukemia	7129	72
3	Prostate	6034	102
4	Lung	12533	181

Table 1: Summary of Datasets

The Number of Genes above Certain Selection Frequencies across 100 Gene Signatures of Size 50 Selected by the Instance weighting, Sample Weighting and Hybrid Weighting algorithm.

Data	Selection Method	Frequency Interval		
		[1,100]	[50,100]	[85,100]
Colon	IW	420	34	12
	SW	350	36	16
	HW	280	39	20
Leukemia	IW	480	26	11
	SW	469	28	14
	HW	456	29	18
Prostate	IW	282	29	13
	SW	262	37	15
	HW	232	41	19
Lung	IW	262	39	18
	SW	246	42	22
	HW	210	48	27

Table 2: Resulting

the first stability of feature selection from gene expression microarray data sets. About stable feature selection framework which defines the perspective sample variance from the stability of feature selection and illustrate that the stability of feature selection under training data variations can be improved by variance reduction techniques. The first contribution of this research is a general framework of weighting to improve the stability of existing feature selection methods. The framework weight each sample in a training set allowing to its influence to the approximation of feature relevance, and then the weighted training set has been provided to a feature selection method.

The second contribution of this research is the margin-based hybrid weighting algorithm developed under the general framework. Hybrid Weighting (HW) for feature selection method is based on the concept of average, hypothesis margin induced by boosting. The weighting criterion term Margin Fraction (MF) is a hybrid along with Sample Weighting (SW). Weighting is assigned by the algorithm to each samples according to the outlying of its local profile of feature relevance (margin vector) compared with other samples.

VII. Conclusion

The research, based on gene expression data sets, has shown that the margin-based hybrid weighting algorithm stability is improving with representative without sacrificing their predictive performance. The results suggest that the general framework of hybrid weighting is a promising approach to improve the stability of feature selection methods for gene selection.

VIII. Future work

This research work can be further extended in the following direction: adding additional weighting algorithm under the framework and investigating their effectiveness on different feature selection methods or using different classification algorithm. Since the hybrid weighting framework is not limited to work with a particular selection method, it is reasonable to expect that this framework could improve the stability of other selection methods. To apply the hybrid weighting framework to other selection methods, take weighting samples as input and consider hybrid weighting in feature evaluation.

References

[1] Abeel, T., Helleputte, T., Peer, Y.V., Dupont, P. and Saeys, Y. "Robust Biomarker Identification for Cancer Diagnosis with Ensemble Feature Selection Methods," Bioinformatics, vol. 26, no. 3, pp. 392-398, 2010.

[2] Alon, U., Barkai, N., Notterman, D.A., Gishdagger, K., Barradagger, S.Y., Mackdagger, D. and Levine, A.J. "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," Proc. Nat'l Academy of Sciences USA, vol. 96, pp. 6745-6750, 199.

[3] Achlioptas, D. and McSherry, F. Fast computation of low rank matrix approximations. In Proceedings of the thirty-third annual ACM symposium on Theory of computing, pages 611-618. ACM, 2001.

[4] Agrawal, R., Imielinski, T. and Swami, A. 1993. Database mining: A performance perspective. IEEE Trans. on Knowledge and Data Engineering 5, 6 (Dec.), 914–925.

[5] Ai, R. and Langley, P. Induction of one-level decision trees. In Proceedings of the Ninth International Conference on Machine Learning, pp.233–240, 1992.

[6] Aizerman, M., Braverman, E. and Rozonoer, L. 1964. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25, 821–837.

[7] Alelyani, S., Liu, H. and Wang, L. The characteristics of the dataset on the selection stability. In Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on, pages 970 977. IEEE, 2011.

[8] Alelyani, S., Tang, J. and Liu, H. Feature selection for clustering: A review. 2013.

[9] Alizadeh, A., Eisen, M.B., Davis, R.E. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403(6769): 503–511.

All rights reserved by *www.ijsrd.com*

[10] Alon, U., Barkai, N., Notterman, Gishdagger, D.A., Barradagger, K.Y., Mackdagge, S.D. and Levine, A. J. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl Acad. Sci. USA 96, 6745–6750.
[11] Alshawabkeh, M. and Kaeli .D. Feature selection for imbalanced data. Poster at Women in Machine Learning (WiML), Dec 2010.
[12] Alshawabkeh, M., Aslam, J.A., Dy, J. and Kaeli, D. Boosting-based feature selection. Poster at Broadening Participation in Data Mining (BPDM) Workshop, co-located with SDM, April 2012.
[13] Alter, O., Brown, P. and Botstein D. Singular value decomposition for genome wide expression data processing and modeling. Proceedings of the National Academy of Sciences, 97(18):10101, 2000.
[14] Anjan Goswami. Department of Computer Science and Engineering" Fast and Exact Out of-Core and Distributed K-Means Clustering 2001.
[15] Appice, A., Ceci, M., Rawles, S. and Flach, P. 2004. Redundant feature elimination for multi-class problems. In Proceedings of the 21st International Conference on Machine learning. 33–40.
[16] Arun, K Pujari. "Data Mining Techniques", Universities Press (India) Limited 2001, ISBN81-7371-3804.
[17] Boulesteix, A.L. and Slawski, M. "Stability and Aggregation of Ranked Gene Lists," Briefings in Bioinformatics, vol. 10, no. 5, pp. 556-568, 2009.
[18] Bartlett, P. 1999. Generalization performance of support vector machines and other pattern classifiers. Advances in Kernel Methods Support Vector Learning, 43–54.
[19] Bauer, E. and Kohavi, R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning, 36:105–139, 1999.
[20] Belkin, M. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15, 1373–1396.

All rights reserved by *www.ijsrd.com*