ENSEMBLE RANDOM PROJECTION FOR MULTI-LABEL CLASSIFICATION WITH APPLICATION TO PROTEIN SUBCELLULAR LOCALIZATION

Shibiao Wan[*], *Man-Wai Mak*[*], *Bai Zhang*[†], *Yue Wang*[‡], *Sun-Yuan Kung*[§]

[*]Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR
[†]Dept. of Pathology, Johns Hopkins Medical Institutions, Maryland, USA
[‡]Bradley Dept. of Electrical and Computer Engineering, Virginia Tech, Virginia, USA
[§]Dept. of Electrical Engineering, Princeton University, New Jersey, USA

ABSTRACT

The curse of dimensionality severely restricts the predictive power of multi-label classification systems. High-dimensional feature vectors may contain redundant or irrelevant information, causing the classification systems suffer from overfitting. To address this problem, this paper proposes a dimensionality-reduction method that applies random projection (RP) to construct an ensemble of multi-label classifiers. The merits of the proposed method are demonstrated through a multi-label protein classification task. Specifically, high-dimensional feature vectors are extracted from protein sequences using the gene ontology (GO) and Swiss-Prot databases. The feature vectors are then projected onto lower-dimensional spaces by random projection matrices whose elements conform to a distribution with zero mean and unit variance. The transformed low-dimensional vectors are classified by an ensemble of one-vs-rest multi-label support vector machine (SVM) classifiers, each corresponding to one of the RP matrices. The scores obtained from the ensemble are then fused for predicting the subcellular localization of proteins. Experimental results suggest that the proposed method can reduce the dimensions by seven folds and impressively improve the classification performance.

Index Terms— Dimension reduction; Random projection; Protein subcellular localization; Multi-label classification; Support vector machines.

1. INTRODUCTION

In machine learning, high-dimensional patterns are often mapped to a lower dimension subspace to avoid the curse of dimensionality [1]. Reducing the dimension of the input patterns can remove redundant or irrelevant information and allow for more reliable classification in the subspace. Actually, dimension reduction are imperative in various domains, such as text categorization [2], image retrieval [3] and gene expression microarray data analysis [4].

In the past three decades, random projection (RP) has emerged as a powerful method for dimension reduction. By using RP, the high dimensional feature vectors are transformed into a much lower-dimensional vectors, which preserve the original geometrical structure and contain less redundant, irrelevant or even detrimental information that might deteriorate classification performance. RP turns out to be a computationally efficient, yet sufficiently accurate method for dimensionality reduction of many high-dimensional datasets [5]. RP is particularly useful for sparse input data in high dimensions as the original data can be reconstructed almost perfectly from data in the lower-dimensional projected space [6]. RP has been widely used in various applications, such as preprocessing text data [7], indexing audio documents [8], processing images [5], learning high-dimensional Gaussian mixture models [9]. Recently, dynamic random projection [10, 11] is successfully applied in biometric template protection and privacy-preserving verification.

Protein subcellular localization is to predict in which part(s) of a cell a protein resides. In recent years, protein subcellular localization has received tremendous attention due to its vitally important roles in elucidating protein functions and identifying drug targets [12, 13]. Computational methods are required to replace time-consuming and laborious wet-lab methods for predicting the subcellular locations of proteins. A predominant scenario in protein subcellular localization prediction is that the dimension of available features is much larger than the number of training samples [14–20]. It is highly expected that the high-dimensional features contain redundant or irrelevant information, causing overfitting and performance degradation.

This paper proposes an ensemble classifier based on random projection (RP) for predicting subcellular localization of multi-label proteins. To make the classifiers more robust, it is necessary to perform random projection of the feature vectors several times due to the random nature of RP. The resulting projected vectors are then presented to an ensemble of one-vs-rest multi-label SVM classifiers. Results demonstrate that the proposed ensemble classifier substantially outperforms the state-of-the-art predictors and that RP is significantly better than the conventional dimension-reduction and feature-selection methods (such as PCA and RFE-SVM [21]) for subcellular localization. This paper also demonstrates that only 3 to 4 applications of RP will be sufficient to construct an ensemble classifier with input dimension that is one-seventh of that of the full-feature classifiers, while at the same time improves the classification performance.

2. RANDOM PROJECTION

The key idea of RP arises from the Johnson-Lindenstrauss lemma [22]:

Given $\epsilon > 0$, a set $\mathcal{X}$ of N points in $\mathcal{R}^T$, and a positive integer $d \geq d_0 = \mathcal{O}(\log N/\epsilon^2)$, there exists $f : \mathcal{R}^T \to \mathcal{R}^d$ such that

$$(1-\epsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\epsilon)\|u-v\|^2$$

for all $u, v \in \mathcal{X}$. A proof can be found in [23].

The lemma suggests that if points in a high-dimensional space are projected onto a randomly selected subspace of suitable dimension, the distances between the points are approximately preserved.

This work was in part supported by The Hong Kong Research Grant Council, Grant No. PolyU5264/09E and HKPolyU Grant No. G-YJ86.

Specifically, the original T-dimensional data is projected onto a d-dimensional ($d \ll T$) subspace, using a $d \times T$ random matrix $\mathbf{R}$ whose columns are unit lengths. A vector $\mathbf{p}_i \in \mathcal{R}^T$ is projected to:

$$\mathbf{p}_i^{RP} = \frac{1}{\sqrt{d}}\mathbf{R}\mathbf{p}_i, \quad (1)$$

where $1/\sqrt{d}$ is a scaling factor, $\mathbf{p}_i^{RP}$ is the projected vector after RP, and $\mathbf{R}$ is a random $d \times T$ matrix.

The choice of the random matrix $\mathbf{R}$ is one of the key points of interest. Practically, as long as the elements $r_{h,j}$ of $\mathbf{R}$ conforms to any distributions with zero mean and unit variance, $\mathbf{R}$ will give a mapping that satisfies the Johnson-Lindenstrauss lemma [5]. For computational simplicity and also the requirement of sparseness, we adopted a simple distribution proposed by Achlioptas [24] for the elements $r_{h,j}$ as follows:

$$r_{h,j} = \sqrt{3} \times \begin{cases} +1 & \text{with probability 1/6,} \\ 0 & \text{with probability 2/3,} \\ -1 & \text{with probability 1/6.} \end{cases} \quad (2)$$

It is easy to verify that Eq. 2 conforms to a distribution with zero mean and unit variance [24] and that $\mathbf{R}$ is sparse.

3. APPLICATION TO PROTEIN SUBCELLULAR LOCALIZATION

3.1. Feature Extraction

Our subcellular localization predictor uses GO information as the features, which has been demonstrated to be superior to other features [25]. Feature extraction involves two steps: (1) retrieval of GO terms; and (2) construction of GO vectors.

(1) Retrieval of GO Terms. For proteins with known accession numbers (ACs), their respective GO terms are retrieved from the Gene Ontology Annotation (GOA) database[1] using the ACs as the searching keys. For a protein without an AC, its amino acid sequence is presented to BLAST [26] to find its homologs, whose top AC is then used as a key to search against the GOA database.

(2) Construction of GO Vectors. Given a dataset, the GO terms of all of its proteins are retrieved by using the procedure described above. Then, the number of distinct GO terms corresponding to the dataset is determined. Suppose T distinct GO terms are found; these GO terms form a GO Euclidean space with T dimensions. For each sequence in the dataset, a GO vector is constructed by matching its GO terms to all of the T GO terms. Term-frequency [27] is used to construct the GO vectors. The term-frequency approach uses the number of occurrences of individual GO terms as the coordinates. Specifically, the GO vector $\mathbf{p}_i$ of the i-th protein is defined as:

$$\mathbf{p}_i = [b_{i,1}, \cdots, b_{i,j}, \cdots, b_{i,T}]^{\mathsf{T}}, b_{i,j} = \begin{cases} f_{i,j} & \text{, GO hit} \\ 0 & \text{, otherwise} \end{cases} \quad (3)$$

where $f_{i,j}$ is the number of occurrences of the j-th GO term (term-frequency) in the i-th protein sequence. The rationale is that the term-frequencies contain important information for classification. Note that $b_{i,j}$'s are analogous to the term-frequencies commonly used in document retrieval.

3.2. Legitimacy of Using RP

As stated in [6], if $\mathbf{R}$ and $\mathbf{p}_i$ satisfy the conditions of the basis pursuit theorem (i.e., both are sparse in a fixed basis), then $\mathbf{p}_i$ can be reconstructed perfectly from a vector that lies in a lower-dimensional space. In fact, the GO vectors and our projected matrix $\mathbf{R}$ satisfy these conditions. As shown in Fig. 1, the number of non-zero entries in the GO vectors tends to be small (i.e. sparse) when compared to the dimension of the GO vectors. Among the 978 proteins in the dataset, a majority of them only have 9 non-zero entries in the 1541-dimensional vectors, and the largest number of non-zero entries is only 45. These statistics suggest that the GO vectors $\mathbf{p}_i$ in Eq. 1 are very sparse.

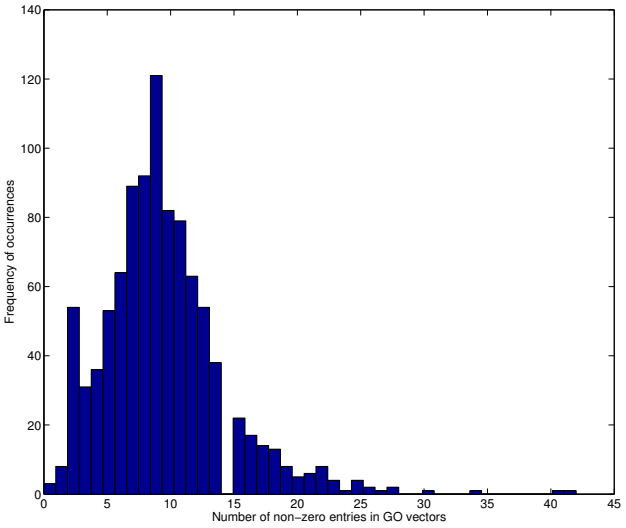

Fig. 1. Histogram illustrating the distribution of the number of non-zero entries (spareness) in the GO vectors with dimensionality 1541. The histogram is plotted up to 45 non-zero entries in the GO vectors because among the 978 proteins in the dataset, none of their GO vectors have more than 45 non-zero entries.

3.3. Ensemble Multi-label Classifier

The projected vectors obtained from Eq. 1 are used for training multi-label one-vs-rest SVMs. Specifically, for an M-class problem (here M is the number of subcellular locations), M independent binary SVMs are trained, one for each class. Denote the GO vector of the i-th query protein as $\mathbf{q}_i$. If the AC of the protein is known, $\mathbf{q}_i$ is created by using the AC; if the AC is unknown, $\mathbf{q}_i$ is created by using the top homologous AC obtained from BLAST. By Eq. 1, we obtained the corresponding projected vector $\mathbf{q}_i^{RP}$. Then, given the i-th query protein $\mathbb{Q}_i$, the score of the m-th SVM is:

$$s_m(\mathbb{Q}_i) = \sum_{r \in \mathcal{S}_m} \alpha_{m,r} y_{m,r} K(\mathbf{p}_r^{RP}, \mathbf{q}_i^{RP}) + b_m \quad (4)$$

where $\mathcal{S}_m$ is the set of support vector indexes corresponding to the m-th SVM, $y_{m,r} \in \{-1, +1\}$ are the class labels, $\alpha_{m,r}$ are the Lagrange multipliers, $K(\cdot,\cdot)$ is a kernel function; here, the linear kernel is used. Note that $\mathbf{p}_r^{RP}$'s in Eq. 4 represents the projected GO training vectors, which may include the projected GO vectors created by using the true AC of the training sequences or their homologous ACs.

Since $\mathbf{R}$ is a random matrix, the scores in Eq. 4 for each application of RP will be different. To address the randomness issue of RP, we construct an ensemble classifier by fusing the scores resulting from several applications of RP, where the ensemble score of the m-th SVM for the i-th query protein is given as follows:

$$s_m^{en}(\mathbb{Q}_i) = \sum_{l=1}^{L} w_l \cdot s_m^{(l)}(\mathbb{Q}_i), \quad (5)$$

where $\sum_{l=1}^{L} w_l = 1$, $s_m^{(l)}(\mathbb{Q}_i)$ represents the score of the m-th SVM for the i-th protein via the l-th application of RP, L is the total num-

[1] http://www.ebi.ac.uk/GOA

ber of applications of RP, and $\{w_l\}_{l=1}^{L}$ are the weights. For simplicity, here we set $w_l = 1/L, l = 1, \ldots, L$. We refer L as 'ensemble size' in the sequel. Unless stated otherwise, the ensemble size was set to 10 in our experiments, i.e., $L = 10$. Note that instead of mapping the original data into an Ld dimensional vector, the ensemble RP projects the data to L d-dimensional vectors.

To predict the subcellular locations of datasets containing both single-label and multi-label proteins, a decision scheme for multi-label SVM classifiers should be used. Unlike the single-label problem where each protein has one predicted label only, a multi-label protein should have more than one predicted labels. In this paper, we used the decision scheme decsribed in mGOASVM [28]. In this scheme, the predicted subcellular location(s) of the i-th query protein are given by:

$$\mathcal{M}^*(\mathbb{Q}_i) = \begin{cases} \bigcup_{m=1}^{M} \{m : s_m^{en}(\mathbb{Q}_i) > 0\}, \text{ where } \exists\, s_m^{en}(\mathbb{Q}_i) > 0\,; \\ \arg\max_{m=1}^{M} s_m^{en}(\mathbb{Q}_i), \text{ otherwise.} \end{cases} \tag{6}$$

For ease of comparison, we refer to the proposed ensemble classifier with this multi-label decision scheme as RP-SVM.

3.4. Datasets and Performance Metrics

A plant dataset [17] was used to evaluate the performance of the proposed predictors. This dataset was created from Swiss-Prot 55.3 and it contains 978 plant proteins distributed in 12 locations. Of the 978 plant proteins, 904 belong to one subcellular locations, 71 to two locations, 3 to three locations and none to four or more locations. The sequence identity was cut off at 25%.

Compared to traditional single-label classification, multi-label classification requires more complicated performance metrics to better reflect the multi-label capabilities of classifiers. These measures include *Accuracy, Precision, Recall, F1-score (F1)*, *Hamming Loss (HL)* [29], overall locative accuracy (*OLA*) [16] and overall actual accuracy (*OAA*) [28]. The last two measures are often used in multi-label subcellular localization prediction. Among all the metrics, *OAA* is the most stringent and objective. This is because if only some (but not all) of the subcellular locations of a query protein are correctly predict, the numerators of the other measures are non-zero, whereas the numerator of *OAA* is 0 (thus contribute nothing to the frequency count). Therefore, we will focus on *OAA*, and unless stated otherwise, the term 'performance' refers to *OAA* thereafter.

4. RESULTS AND DISCUSSIONS

4.1. Performance of Ensemble Random Projection

Fig. 2 shows the performance of ensemble RP-SVM for different feature dimensions based on leave-one-out cross-validation (LOOCV). The cyan dotted line represents the performance of mGOASVM [28]. In other words, it represents the performance achieved by the full features, which are referred to as full-dim SVM in the figure legend. The dimensionality of the original feature vectors is 1541. As can be seen, for dimensions between 200 and 1400, the performance of RP-SVM is better than that of mGOASVM, which demonstrates that RP can boost the classification performance even the dimension is only one-seventh (200/1400) of that of the original one. This suggests that the original feature vectors really have irrelevant or redundant information.

Fig. 2 also shows the performance statistics of RP-SVM based on LOOCV at different feature dimensions, when the ensemble size (L in Eq. 5) is fixed to 1, which we refer to as 1-RP-SVM. We created ten 1-RP-SVM classifiers, each with a different RP matrix. The result shows that the mean accuracy of the ten 1-RP-SVM is lower than that of mGOASVM when the projected dimension d is below

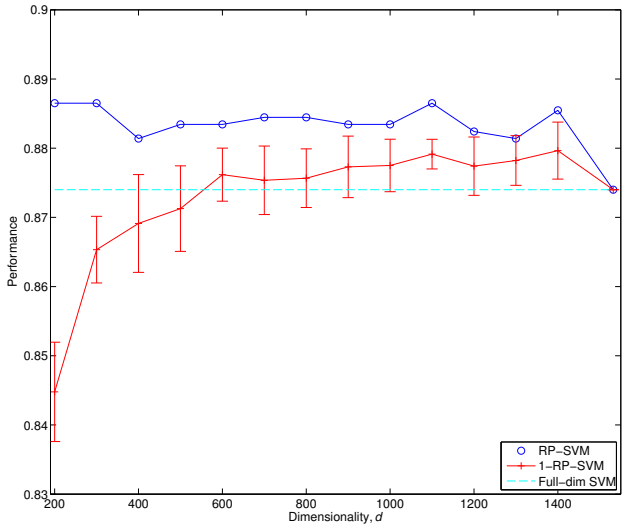

Fig. 2. Performance of ensemble RP-SVM and 1-RP-SVM (with standard deviation error bars) for different feature dimensions based on leave-one-out cross-validation (LOOCV) on the plant dataset. The cyan dotted lines represent the performance of mGOASVM [28] (using full-dimensional features, referred as to full-dim SVM). *1-RP-SVM*: RP-SVM with an ensemble size of 1.

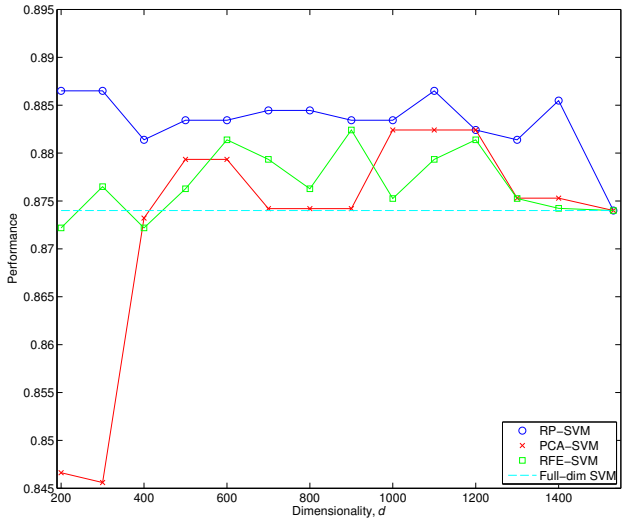

Fig. 3. Comparing ensemble random projection with other dimension-reduction methods at different projected dimensions based on LOOCV. The cyan dotted lines represent the performance of mGOASVM (full-dim SVM) .

600, suggesting that there is a lower limit for the projected dimension. Moreover, the error bars in Fig. 2 show that when the projected dimension is below 1000, the performance of some of the 1-RP-SVM classifiers is poorer than that of mGOASVM. This suggests that the performance can be poorer than that using the full-features when a single RP is used for the projection, especially at low projection dimension.

Fig. 2 shows that the ensemble RP-SVM performs significantly better than 1-RP-SVM for a wide range of dimensionality. This suggests that the ensemble scheme can boost the performance to a level that is higher than any of the individual RPs.

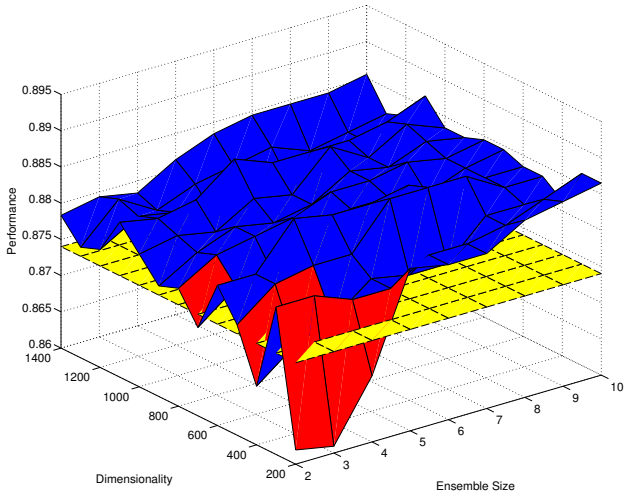

Fig. 4. Performance of RP-SVM at different projected dimensions and different ensemble sizes of random projection. The yellow dotted plane represents the accuracy of mGOASVM, which is a constant for all projected dimensions and ensemble size. The mesh with blue (red) surfaces represent the projected dimensions and ensemble sizes at which the ensemble RP-SVM performs better (poorer) than mGOASVM. *Ensemble Size*: Number of times of random projection for ensemble.

4.2. Comparing with Other Dimension-Reduction Methods

Fig. 3 compares RP-SVM with other dimension-reduction methods based on LOOCV. Here, PCA-SVM and RFE-SVM mean replacing RP with principal component analysis (PCA) and recursive feature elimination [21]. As can be seen, RP-SVM performs the best for a wide range of projected dimensions, while RFE-SVM and PCA-SVM perform poorly when the dimension is reduced to 200 (out of 1541). Only when the dimension is around 1200, all the three methods perform equally. This suggests that RP-SVM is better than PCA-SVM and RFE-SVM for reducing the dimension of GO vectors. This is reasonable because the GO vectors are very sparse, which is more suitable for RP than for PCA and RFE.

4.3. Effect of Dimensions and Ensemble Size

As individual RP cannot guarantee good performance, it is reasonable to ask: at least how many times of RP can guarantee that the performance of the ensemble classifier is equivalent to, or even better than that of the one without RP (i.e., mGOASVM)? Fig. 4 shows the performance of RP-SVM for different dimensions and different ensemble sizes of RPs on the plant dateset. The blue/red areas represent the condition under which RP-SVM performs better/worse than mGOASVM. The yellow dotted plane represents the performance of mGOASVM. As can be seen, for dimensionality from 300 to 1400, RP-SVM with at least 4 applications of RP can outperform mGOASVM; for dimensionality 200, we need at least 5 applications of RP to obtain a performance better than mGOASVM. These results suggest that the proposed RP-SVM is very robust because only 4 applications of RP will be sufficient to achieve good performance.

4.4. Comparing with State-of-the-Art Predictors

Table 1 compares the performance of RP-SVM against several state-of-the-art multi-label predictors on the plant dataset. All of the predictors use the information of GO terms as features. From the classification perspective, Plant-mPLoc [16] uses an ensemble OET-KNN (optimized evidence-theoretic K-nearest neighbors) classifier; iLoc-Plant [17] uses a multi-label KNN classifier; mGOASVM [28] uses a multi-label SVM classifier.

Table 1. Comparing the performance of the proposed RP-SVM with state-of-the-art multi-label predictors on the plant dataset. "–" means the corresponding references do not provide the related metrics. *SCL*: subcellular location, including cell membrane (Mem), cell wall (Wal), Chloroplast (Chl), Cytoplasm (Cyt), Endoplasmic Reticulum (ER), Extracellular (Ext), Golgi apparatus (Gol), Mitochondrion (Mit), Nucleus (Nuc), Peroxisome (Per), Plastid (Pla) and Vacuole (Vac). ^{a}, ^{b} and ^{c} are from Refs. [16, 17, 28]. The p-value between the OAA of RP-SVM and mGOASVM is 2.021×10^{-4}.

SCL	LOOCV Locative Accuracy (LA)			
	Plant-mPLoc^a	iLoc-Plant^b	mGOASVM^c	RP-SVM
Mem	0.429	0.696	0.946	0.964
Wal	0.250	0.594	0.844	0.906
Chl	0.867	0.881	0.951	0.993
Cyt	0.396	0.626	0.956	0.945
ER	0.405	0.500	0.905	0.929
Ext	0.136	0.091	1.000	0.955
Gol	0.286	0.762	0.905	0.905
Mit	0.760	0.747	1.000	1.000
Nuc	0.895	0.921	0.993	0.974
Per	0.667	0.286	1.000	1.000
Pla	0.103	0.179	1.000	0.949
Vac	0.500	0.538	0.942	0.962
OLA	0.637	0.717	0.962	**0.971**
OAA	–	0.681	0.874	**0.887**
Accuracy	–	–	0.926	**0.938**
Precision	–	–	0.933	**0.946**
Recall	–	–	0.968	**0.979**
F1	–	–	0.942	**0.954**
HL	–	–	0.013	**0.011**

As shown in Table 1, RP-SVM performs significantly better than Plant-mPLoc and iLoc-Plant. Both the *OLA* and *OAA* of RP-SVM are more than 20% (absolute) higher than iLoc-Plant. When comparing with mGOASVM, the *OAA* and *OLA* of RP-SVM are also higher than that of mGOASVM, respectively. In terms of *Accuracy, Precision, Recall, F1* and *HL*, RP-SVM perform better than mGOASVM. The results suggest that the proposed RP-SVM performs better than the state-of-the-art classifiers. The individual locative accuracies of RP-SVM are remarkably higher than that of Plant-mPLoc and iLoc-Plant, and are higher than or comparable to mGOASVM. The p-value between the OAA of RP-SVM and mGOASVM is 2.021×10^{-4}, which suggests that the performance of RP-SVM is significantly better than that of mGOASVM.

5. CONCLUSIONS

This paper proposes an ensemble multi-label classifier constructed from a dimension-reduction method based on random projection for protein subcellular localization prediction. Given a query protein, a GO-based feature vector is constructed by exploiting the information in the gene ontology annotation database. The GO-vector is projected onto a much lower-dimensional space by random matrices whose elements conform to Achlioptas' distribution, which are subsequently classified by multi-label SVMs classifiers. By fusing SVM scores obtained via several applications of individual RP, a robust multi-label ensemble classifier can be obtained. It was found that an ensemble size of 3 or 4 will be sufficient to achieve good performance and reduce the feature dimensions by as many as seven folds.

6. REFERENCES

[1] R. Lotlikar and R. Kothari, "Adaptive linear dimensionality reduction for classification," *Pattern Recognition*, vol. 33, no. 2, pp. 185–194, 2000.

[2] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *The Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[3] D. L. Swets and J. J. Weng, "Efficient content-based image retrieval using automatic feature selection," in *Proceedings of International Symposium on Computer Vision*. IEEE, 1995, pp. 85–90.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[5] E. Bingham and H. Mannila, "Random projection in dimension reduction: Applications to image and text data," in *the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*, 2001, pp. 245–250.

[6] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[7] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," in *Proceedings of the 17th ACM Symposium on the Principles of Database Systems*, 1998, pp. 159–168.

[8] M. Kurimo, "Indexing audio documents by using latent semantic analysis and SOM," in *Kohonen Maps*. 1999, pp. 363–374, Elsevier.

[9] S. Dasgupta, "Learning mixtures of Gaussians," in *40th Annual IEEE Symposium on Foundations of Computer Science,*, 1999, pp. 634–644.

[10] B. Yang, D. Hartung, K. Simoens, and C. Busch, "Dynamic random projection for biometric template protection," in *2010 Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, 2010, pp. 1–7.

[11] Y. Wang and K. N. Plataniotis, "An analysis of random projection for changeable and privacy-preserving biometric verification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 5, pp. 1280–1293, 2010.

[12] G. Lubec, L Afjehi-Sadat, J. W. Yang, and J. P. John, "Searching for hypothetical proteins: theory and practice based upon original data and literature," *Prog. Neurobiol*, vol. 77, pp. 90–127, 2005.

[13] S. Wan, M. W. Mak, and S. Y. Kung, "Semantic similarity over gene ontology for multi-label protein subcellular localization," *Engineering*, vol. 5, pp. 68–72, 2013.

[14] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *J. Mol. Biol.*, vol. 238, pp. 54–61, 1994.

[15] S. Wan, M. W. Mak, and S. Y. Kung, "Protein subcellular localization prediction based on profile alignment and Gene Ontology," in *2011 IEEE International Workshop on Machine Learning for Signal Processing (MLSP'11)*, Sept 2011, pp. 1–6.

[16] K. C. Chou and H. B. Shen, "Plant-mPLoc: A top-down strategy to augment the power for predicting plant protein subcellular localization," *PLoS ONE*, vol. 5, pp. e11335, 2010.

[17] Z. C. Wu, X. Xiao, and K. C. Chou, "iLoc-Plant: A multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites," *Molecular BioSystems*, vol. 7, pp. 3287–3297, 2011.

[18] S. Wan, M. W. Mak, and S. Y. Kung, "GOASVM: Protein subcellular localization prediction based on gene ontology annotation and SVM," in *2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'12)*, 2012, pp. 2229–2232.

[19] S. Mei, "Multi-label multi-kernel transfer learning for human protein subcellular localization," *PLoS ONE*, vol. 7, no. 6, pp. e37716, 2012.

[20] S. Wan, M. W. Mak, and S. Y. Kung, "Adaptive thresholding for multi-label SVM classification with application to protein subcellular localization prediction," in *2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'13)*, 2013, pp. 3547–3551.

[21] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[22] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," in *Conference in Modern Analysis and Probability*, 1984, pp. 599–608.

[23] P. Frankl and H. Maehara., "The Johnson-Lindenstrauss lemma and the sphericity of some graphs," *Journal of Combinatorial Theory, Series B*, vol. 44, pp. 355–362, 1988.

[24] D. Achlioptas, "Database-friendly random projections: Johnson-Lindenstrauss with binary coins," *Journal of Computer and Systems Sciences*, vol. 66, pp. 671–687, 2003.

[25] K. C. Chou and H. B. Shen, "Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers," *J. of Proteome Research*, vol. 5, pp. 1888–1897, 2006.

[26] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, pp. 3389–3402, 1997.

[27] S. Wan, M. W. Mak, and S. Y. Kung, "GOASVM: A subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 323, pp. 40–48, 2013.

[28] S. Wan, M. W. Mak, and S. Y. Kung, "mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines," *BMC Bioinformatics*, vol. 13, pp. 290, 2012.

[29] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hullermeier, "On label dependence and loss minimization in multi-label classification," *Machine Learning*, vol. 88, no. 1-2, pp. 5–45, 2012.