Random Projections of Fischer Linear Discriminant Classifier for Multi-Class Classification

Ishank Arora[a*], Anant Dadu[b*], Mridula Verma[c], K. K. Shukla[d]
Department of Computer Science and Engineering
Indian Institute of Technology (BHU), Varanasi, India
Email: ishank.arora.cse14@iitbhu.ac.in[a], anant.dadu.cse14@iitbhu.ac.in[b],
mridula.rs.cse13@iitbhu.ac.in[c], kkshukla.cse@iitbhu.ac.in[d]
*These authors contributed equally to this work.

***Abstract*—Ensembling classifiers has been an effective technique for improving performance and generalizability of classification tasks. In a recent research direction, the ensemble of the random projections is being utilized as an effective regularization technique with linear discriminant classifiers. However the framework has only been designed for binary classifiers. In this paper we extend the idea for the multiclass classifiers, which directly improves the applicability of the framework to a broader class of problems. We performed experiments with multiple high-dimensional benchmark datasets, and compare the performance of our framework with other state-of-the-art methods for multi-class classification. We also extend the theoretical error bounds for misclassification to provide a theoretical analysis. Results demonstrate the efficacy of our methodology.**

Keywords—Random projections, Ensemble learning, Fischer Linear Discriminant, Multi-class classification.

I. Introduction

In today's age, machine learning is an incredibly powerful tool that enables us to crunch petabytes of data and make sense of all the complications that lie within it. It is becoming increasingly ubiquitous with more and more applications in places where we cannot even think of, such as genetics, anomaly detection, and speech recognition to name a few.

Ensembling is one of the most successful techniques used these days to predict the output for both classification as well as regression problems. Algorithms like Random forests [1], Random subspaces [2] and Bagging [3] use different ensemble methods, combining several weak learners to build a strong learning algorithm. Particularly, ensembling of classifiers using random projections proved to be very effective in the papers [4], [5], [6], but none of them was able to do a theoretical analysis.

Categorizing the output in two classes is termed binary classification and when output contains more than two classes it is called multi-class classification. Multi-class classification has been of intrigue for some time now as it finds multiple applications in image processing and medical research, and methods have been proposed to implement it using pairwise coupling [7] as well as using association rules [8]. For instance, it can be used to classify a flower as one of the over 300,000 known flowering plants [9], classify a whistled tune from 30 million recorded songs [10] or recommend physical therapy for the injured [11]. Among the algorithms to convert a binary classification algorithm to multi-class, the common ones are one-vs-rest (OvR) [12] in which real valued confidence score for each class is obtained and the label which has highest confidence score is predicted as the output, and the one-vs-one (OvO) [13] in which we build classifiers using all pairs of multiple classes and the class with the maximum frequency for a given test sample is chosen as the predicted output.

In this paper, the base classifier in consideration is the Fisher Linear Discriminant (FLD) [14] classifier, considered an optimal choice for the high-dimensional domain. FLD is a generative model classifier which searches the most optimal decision boundary among the classes. It projects high-dimensional data onto one dimensional space and maximizes the distance between means of the the two labels, thus minimizing the variance within each label. FLD classifier has proved to be very successful in speech recognition [15]. There have been previous works [16], [17] which prove the application of the classifier in face recognition as well, as well as in the multi-class text categorization task in [18]. The classifier attained impressive results on the WebKB4 dataset having 4 different categories, TDT2 dataset with 96 classes and the Industry Sector dataset.

The ensembling technique in consideration is to implement the FLD classifier by introducing randomly projected matrices. Binary classification using the aforementioned ensemble, along with theoretical guarantees for regularization and boundedness of generalization error for datasets with more features than samples, has been presented in [19], with very promising results on high dimensional datasets like Dorothea (1000000 dimensions).

We extend the ensemble to perform multi-class classification, the condition still being that the number of dimensions be greater than the number of samples. Our ensemble projects the high dimensional data on a randomly generated low dimensional space and maximizes the distance between the means of different classes in the projected space. The OvR method has been implemented to make randomly projected FLD classifier work for multi class classification. The proposed method was tested on three datasets of the bioinformatics and textual domain, settings in which fewer observations than dimensions are the norm, and results obtained prove that the proposed

ensemble is competitive with the state of the art benchmarks. We also extend the generalization error bounds for the multi class classification using the theoretical proofs from [19] and [20].

The organization of our paper is as follows: Section II provides some background details and the description of the randomly projected FLD classifier in binary class domain. Our extension for multi-class classification, including the theoretical error bounds, is presented in Section III. Section IV presents our experimental observations on various publicly available datasets. We finally conclude the paper and our future research endeavor in Section V.

II. Randomly Projected Fischer Linear Discriminant for Binary Classification

Let us consider the binary classification problem with the training data comprising of N $i.i.d.$ samples each having d features, $T_N = \{(x_i, y_i) : x_i \in R^d, y_i \in \{0,1\}\}_{i=1}^N$. The formulation for Fischer Linear Discriminant Classifier is defined by indicator function $1(.)$ which predicts 1 if the value of argument comes out to be true and 0 otherwise is given by:

$$\hat{h}(x_q) := \mathbf{1}\left\{(\hat{\mu_1} - \hat{\mu_0})^T \hat{\Sigma}^{-1}\left(x_q - \frac{\hat{\mu_0} + \hat{\mu_1}}{2}\right) > 0\right\}$$

where $\hat{\mu}_0, \hat{\mu}_1$ are the maximum likelihood estimates of the corresponding class-conditional means for negative and positive samples respectively and $\hat{\Sigma}$ is the shared covariance matrix.

The setting in consideration, as mentioned previously, is that the number of features outnumbers the number of samples, or $d >> N$, resulting in $\hat{\Sigma}$ being singular, and thus non-invertible. So the d dimensional space is projected into k dimensional space using the $i.i.d$ random matrix $R \in M_{k*d}$ [21]. Ensembling is performed by generating M such random matrices, hence M FLD classifiers and averaging their outputs to predict the class based on the average linear decisions of these base classifiers. This combination rule is called *voting* in the ensemble literature. As expectation and the projection operator follow the linearity principle, the decision rule for ensembled randomly projected FLD classifier is given by:

$$\hat{h}_e(x_q) := \mathbf{1}\left\{\frac{1}{M}\sum_{i=1}^{M}\left((\hat{\mu_1} - \hat{\mu_0})^T R_i^T (R_i \hat{\Sigma} R_i^T)^{-1} R_i \right.\right.$$
$$\left.\left. ...\left(x_q - \frac{\hat{\mu_0} + \hat{\mu_1}}{2}\right)\right) > 0\right\}$$

A. Generalization Error Bound Theorem

Theorem 1 [19] Let x_q is the given query with the estimated model covariance matrix $\hat{S}^{-1}$ and the true class-conditional covariance Σ, such that Σ is full rank and $[Pr_{y_q=y}] = \pi_y$. Let ρ be the rank of maximum likelihood estimate of the covariance matrix and let $k < \rho - 1$ be a positive integer. N is the size of training set and N_0, N_1 are the number of samples belonging to each class respectively. Then for $\delta \in (0,1)$ the generalization error upper bound [19] with probability atleast $1 - \delta$ is given by:

$$\Pr_{x_q, y_q}(\hat{h}_e(x_q) \neq y_q) \leq \sum_{y=0}^{1} \pi_y \Phi\left(-\left[g\left(\overline{\kappa}\left(\sqrt{2log\frac{5}{\delta}}\right)\right)\right.\right.$$
$$... * \left[\sqrt{\|\Sigma^{-\frac{1}{2}}(\hat{\mu}_1 - \hat{\mu}_0)\|^2 + \frac{dN}{N_0 N_1}} - \sqrt{\frac{2N}{N_0 N_1} log\frac{5}{\delta}}\right]_+$$
$$\left.\left. ... - \sqrt{\frac{d}{N_y}}\left(1 + \sqrt{\frac{2}{d} log\frac{5}{\delta}}\right)\right]\right)$$

where Φ is the c.d.f. of the standard Gaussian, $\overline{\kappa}(\epsilon)$ is a high probability upper bound on the condition number of $\Sigma\hat{S}^{-1}$ and $g(.)$ is the function $g(a) := \frac{\sqrt{a}}{1+a}$.

Taking the random projection is particularly useful as compared to a regularized FLD when considered from a computational point of view. Further, parallel implementation is possible, both for training and classification, as the individual ensembles are executed independent of each other.

III. Multi-Class Randomly Projected Fischer Linear Discriminant

One of the main contributions of this paper is its extension to classify multi-class datasets. The one vs rest (OvR) technique has been used for the same, in which a classifier is trained for each and every class by taking that particular class as positive and all other classes as negative. Each class is assigned a confidence score for a given test sample and the class which has the highest confidence score is chosen as the predicted class for that sample. Let $C \in \{0, 1, ..., K\}$ be the set of $K + 1$ different classes in which the data has been partitioned so each sample can have any one value from the set C. The decision rule for this proposed hypothesis is given by:

$$\hat{h}_{ens}(x_q) := \underset{y \epsilon C}{\operatorname{argmax}}\left\{\frac{1}{M}\sum_{i=1}^{M}\left((\hat{\mu}_y - \hat{\mu}_{\neg y})^T\ R_i^T\right.\right.$$
$$\left.\left. ...(R_i \hat{\Sigma} R_i^T)^{-1} R_i\left(x_q - \frac{\hat{\mu}_y + \hat{\mu}_{\neg y}}{2}\right)\right) > 0\right\}$$

A. Multi-Class Error Bound Extension

In the multi-class case, the probabilistic decision rule is given by:

$$\hat{h}_{ens}(x_q) = j \iff j = \underset{i}{\operatorname{argmax}}\{\Pr_y(i|x_q)\}\, \forall y, i, j \in C$$

We can consider the correct class to be 0 and find the generalization error with respect to this class and then similarly extend this for all the classes. Hence,

$$\hat{h}_{ens}(x_q) = 0 \iff \bigwedge_{i \neq 0}\{\Pr_y(0|x_q) \geq \Pr_y(i|x_q)\}$$

So, the misclassification for class 0, using De Morgan's Laws on the above equation, is given by

$$\hat{h}_{ens}(x_q) \neq 0 \iff \bigvee_{i \neq 0}\{\Pr_y(i|x_q) > \Pr_y(0|x_q)\}$$

TABLE I
DATASETS

Name	No. of samples	No. of features	No. of classes	$\rho/2$	Max. Accuracy (%)
News20	15935	62061	20	7966	88.13
Leukemia	215	12558	7	107	94.64
Cancer Multi-A	63	5565	4	31	97.5

TABLE II
COMPARISON OF VARIOUS CLASSIFIERS

	Accuracy (%)													
	News20					Leukemia					Cancer Multi-A			
	Number of training samples					Number of training samples					Number of training samples			
Classifier	2000	4000	6000	8000	15935	100	125	150	175	215	30	40	50	63
Decision Tree	50.23	53.24	56.14	58.22	59.30	58.92	64.28	69.64	81.25	79.46	80.0	85.0	82.5	95.0
Random Forest	59.40	62.58	67.59	67.56	70.22	69.64	75.89	79.46	75.89	89.28	65.0	85.0	90.0	97.5
AdaBoost	39.49	46.03	47.10	47.50	48.08	44.64	52.67	60.71	66.96	65.17	57.5	67.5	65.0	60.0
RP-FLD($M = 10, k = \rho/2$)	57.28	63.46	70.74	77.01	88.13	65.18	76.79	86.61	91.96	94.64	87.5	95.0	92.5	97.5

Considering the probabilities, since if $A = B \implies Pr(A) = Pr(B)$ we have,

$$\Pr_{x_q}[\hat{h}_{ens}(x_q) \neq 0] = \Pr_{x_q}\left[\bigvee_{i\neq 0}\{\Pr_y(i|x_q) > \Pr_y(0|x_q)\}\right] \leq \sum_{i=1}^{K} \Pr_{x_q}\left\{\frac{\Pr_y(i|x_q)}{\Pr_y(0|x_q)} > 1\right\}$$

Now for $Pr_{x_q}[\hat{h}_{ens}(x_q) \neq 0]$ we get the bounded error by summing the binary classification error over the classes $\{1, ..., K\}$. Thus using Theorem 1, we get the result of bounded misclassification error for multi-class scenarios,

$$\Pr_{x_q, y_q}(\hat{h}_{ens}(x_q) \neq y_q) \leq \sum_{y=0}^{K}\sum_{i\neq y}\pi_y \Phi\Bigg(-\Bigg[g\left(\overline{\kappa}\left(\sqrt{2log\frac{5}{\delta}}\right)\right) ... * \left[\sqrt{\|\Sigma^{-\frac{1}{2}}(\hat{\mu}_i - \hat{\mu}_y)\|^2 + \frac{dN}{N_i N_y}} - \sqrt{\frac{2N}{N_y N_i} log\frac{5}{\delta}}\right]_+ ... - \sqrt{\frac{d}{N_y}}\left(1 + \sqrt{\frac{2}{d} log\frac{5}{\delta}}\right)\Bigg]\Bigg)$$

IV. EXPERIMENTAL RESULTS

To validate the prediction performance of our proposed ensemble, experiments were conducted to classify the output for three publicly available multiclass datasets, each characterized by having more features than samples. The first dataset, News20 [22] from the textual domain is a collection of news documents partitioned across 20 differents newsgroups, each corresponding to a different topic. The Leukemia (Stjude) [23] dataset classifies the types of pediatric acute lymphoblastic leukemia across six diagnostic groups and one group of outliers. The third dataset, Cancer Multi-A (Diffuse Large B-cell lymphoma-A) [24] is from the bioinformatics domain. The experiments were performed in Python 2.7.10 on a 64-bit OS, x64-based processor, Windows 10 machine, Intel® Core™ i7-5500U CPU @ 2.40 GHz, having 8.00 GB RAM. For the News20 dataset, the classification was performed using 64-bit Intel® Xeon Processor E5-2420 v2 @ 2.20GHz server, running Arch-Linux having 192 GB RAM.

The characteristics of the datasets along with the results obtained are presented in Table I. In the first set of experiments, classification using the randomly projected FLD was performed by varying the number of classifiers while keeping the number of dimensions of the projected space constant. As can be observed from the graphs in fig. 2, increasing the number of classifiers causes the accuracy of classification to reach a saturation, and a further increase in the number does not bring about a significant rise in accuracy, but leads to a linear increase in the computational overhead. So only a marginal increase in the degree of accuracy is seen as the number of classifiers, M, varies from 10 to 20 in the Leukemia dataset, and from 20 to 30 in the Cancer Multi-A dataset.

Another set of classification experiments was conducted by varying the number of projection dimensions while keeping the number of classifiers constant, the results of which have been plotted in fig. 3. It was observed that the best results were obtained when the projection dimension had a value in the vicinity of $\rho/2$, where ρ is the rank of the shared covariance matrix, $\hat{\Sigma}$. The choice of the projection dimension, k, works as a regularizer implementing feature selection, and $k = \rho/2$ proves to be an optimal scheme. Choosing a large value of k causes highly correlated features to be taken into consideration, while a very small value leads to loss of vital information.

For comparisons with the state of the art, in the third set of

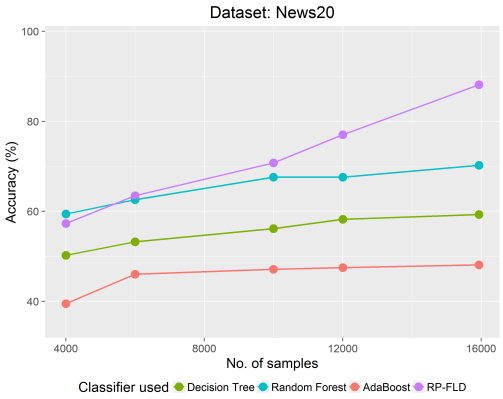

Fig. 1. Comparison of various classifiers with the ensemble of Randomly Projected FLD classifier by varying the number of training samples on the News20 dataset.

experiments, the datasets were classified using the widely used Decision Tree, Random Forest and the AdaBoost classifiers, available in the Scikit-learn [25] module in Python 2.7, by varying the number of training samples for the three datasets, as presented in fig. 1 for News20, and in fig. 4 for Leukemia and Cancer Multi-A. As can be observed from Table II, the proposed ensemble, when trained with a sufficient number of classifiers and an optimal projection dimension of around $\rho/2$, outperforms the existing algorithms in terms of the degree of accuracy. As for each instance of random projection, a classifier is trained considering each class as the correct output, the time complexity of the proposed ensemble varies linearly with the number of classes in the dataset under consideration. Thus a larger computational overhead is incurred as compared to the other classifiers. So optimal values for the parameters, k and M can be chosen so as to best balance the speed and accuracy of classification.

V. Conclusion

In this paper we propose a multiclass classification framework for the ensemble of randomly projected Fisher Linear Discriminant classifiers. Our theory guarantees that the sum of misclassification errors for all classes remains bounded for optimal choices of the projection dimension. We also demonstrate the practical applicability of the framework with extensive experimentation on multiple benchmark high-dimensional datasets. Consistently better performances than other known classifiers presents a very strong case for its practical usage. In future, we intend to extend our framework for multimodal classification, wherein the dataset is generated from separate model parts which may overlap or may have insufficient coverage of the data space. Furthermore, deriving high probability guarantees on the performance of the finite ensemble also presents a challenging task as the rank deficiency of the shared covariance matrix is difficult to handle.

References

[1] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324

[2] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998. [Online]. Available: http://dx.doi.org/10.1109/34.709601

[3] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996. [Online]. Available: http://dx.doi.org/10.1023/A:1018054314350

[4] *Face recognition experiments with random projection*, vol. 5779, 2005. [Online]. Available: http://dx.doi.org/10.1117/12.605553

[5] R. Folgieri, "Ensembles based on random projection for gene expression data analysis," Ph.D. dissertation, University of Milano, 2007.

[6] A. Schclar and L. Rokach, *Enterprise Information Systems: 11th International Conference, ICEIS 2009, Milan, Italy, May 6-10, 2009. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, ch. Random Projection Ensemble Classifiers, pp. 309–316. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-01347-8_26

[7] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, Dec. 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1005332.1016791

[8] F. A. Thabtah, P. I. Cowling, and Y. Peng, "Mcar: multi-class classification based on association rule." in *AICCSA*. IEEE Computer Society, 2005, p. 33. [Online]. Available: http://dblp.uni-trier.de/db/conf/aiccsa/aiccsa2005.html#ThabtahCP05

[9] A. J. Paton, N. Brummitt, R. Govaerts, K. Harman, S. Hinchcliffe, B. Allkin, and E. N. Lughadha, "Towards target 1 of the global strategy for plant conservation: a working list of all known plant speciesprogress and prospects," *Taxon*, vol. 57, no. 2, pp. 602–611, 2008.

[10] D. Eck, "Personal communication from Google music expert Douglas Eck," 2013.

[11] J. Zhang, D. Gross, and O. R. Zaïane, *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2013 International Workshops: DMApps, DANTH, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD, Australia, April 14-17, 2013, Revised Selected Papers*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, ch. On the Application of Multi-class Classification in Physical Therapy Recommendation, pp. 143–154. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-40319-4_13

[12] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, Dec. 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1005332.1005336

[13] J. Fürnkranz, "Round robin classification," *J. Mach. Learn. Res.*, vol. 2, pp. 721–747, Mar. 2002. [Online]. Available: http://dx.doi.org/10.1162/153244302320884605

[14] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[15] E. Alexandre-Cortizo, M. Rosa-Zurera, and F. Lopez-Ferreras, "Application of fisher linear discriminant analysis to speech/music classification," in *Computer as a Tool, 2005. EUROCON 2005. The International Conference on*, vol. 2. IEEE, 2005, pp. 1666–1669.

[16] C. Liu and H. Wechsler, "Enhanced fisher linear discriminant models for face recognition," in *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, vol. 2. IEEE, 1998, pp. 1368–1372.

[17] C. Xiang, X. Fan, and T. H. Lee, "Face recognition using recursive fisher linear discriminant," *Image Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2097–2105, 2006.

[18] T. Li, S. Zhu, and M. Ogihara, "Using discriminant analysis for multi-class classification: an experimental investigation," *Knowledge and information systems*, vol. 10, no. 4, pp. 453–472, 2006.

[19] R. J. Durrant and A. Kabán, "Random projections as regularizers: Learning a linear discriminant from fewer observations than dimensions," *Mach. Learn.*, vol. 99, no. 2, pp. 257–286, May 2015. [Online]. Available: http://dx.doi.org/10.1007/s10994-014-5466-8

[20] R. J. Durrant and A. Kaban, "Compressed fisher linear discriminant analysis: classification of randomly projected data," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, 2010, pp. 1119–1128. [Online]. Available: http://doi.acm.org/10.1145/1835804.1835945

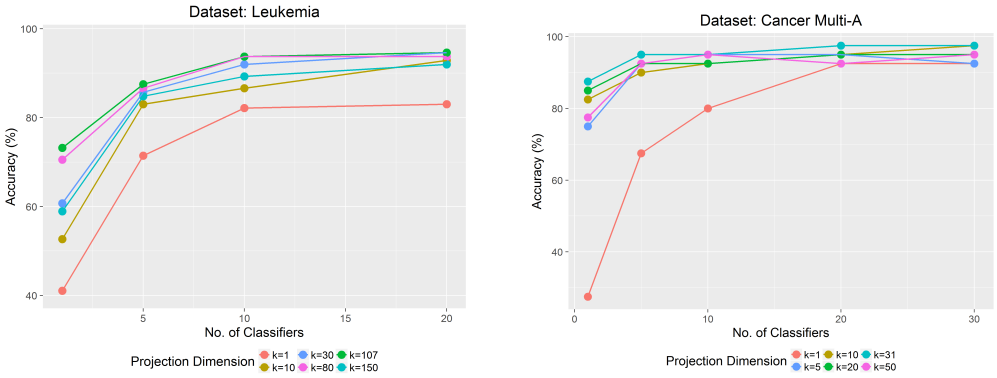

Fig. 2. Effect of variation of the number of classifiers on the Leukemia and Cancer Multi-A datasets.

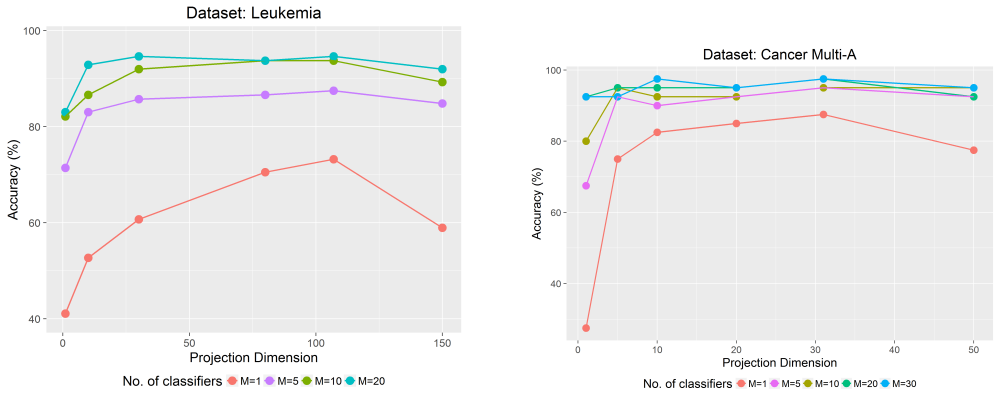

Fig. 3. Effect of variation of the number of projection dimensions on the Leukemia and Cancer Multi-A datasets.

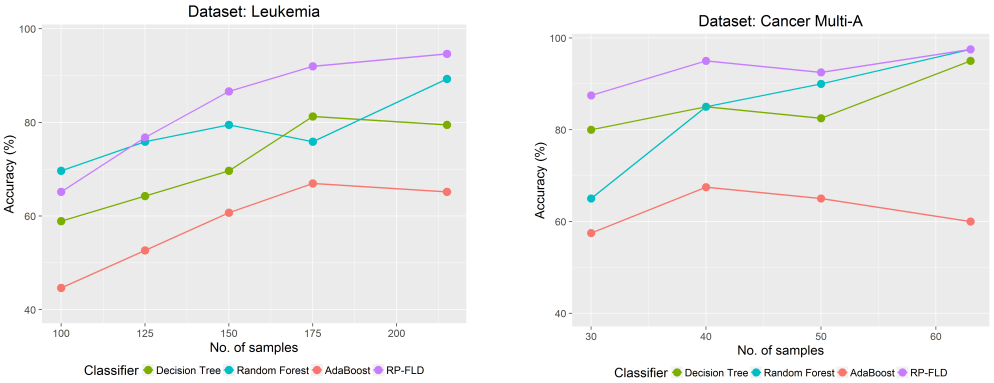

Fig. 4. Comparison of various classifiers with the ensemble of Randomly Projected FLD classifier by varying the number of training samples on the Leukemia and Cancer Multi-A datasets.

[21] T. L. Marzetta, G. H. Tucci, and S. H. Simon, "A random matrix-theoretic approach to handling singular covariance estimates," *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6256–6271, Sept 2011.

[22] K. Lang, "Newsweeder: Learning to filter netnews," in *in Proceedings of the 12th International Machine Learning Conference (ML95*, 1995.

[23] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, W. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.-H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133 – 143, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1535610802000326

[24] Y. Hoshida, J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Subclass mapping: identifying common subtypes in independent disease data sets." *PLoS ONE*, vol. 2, no. 11, 2007. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0001195

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.