NAZARBAYEV UNIVERSITY

Term Project

BENG 146 Programming for Engineers

Level 4 Has been Completely Implemented

Group 14

Group members:

Abylaikhan Bozzhigitov 100 “Second Student in the Presentation”, 201468408

Marzhan Bekbola 100 “Third Student in the Presentation”, 201466501

Meruyert Zhuvandykova 100 ”First and Sixth Student in the Presentation”, 201450089

Miras Naizakarayev 75 “Fifth Student in the Presentation”, 201566188

Mukhtar Turarbek 100 “Fourth Student in the Presentation”, 201497773

Abstract

This project comprises data samples for gene expressions of cancer. These samples are divided into two groups of people. One group is patients with good prognosis, who are described as the class 1; and the another group is patients with bad prognosis, who are assumed as class 0. Therefore, in order to analyze gene samples of future patients, and give them appropriate classification, it is necessary to construct mathematical model by using different classifiers. However, it is also important to identify the probability of possible deviation. Therefore, this project, firstly, will show used classifiers such as LDA, DLDA, QDA, G13, SEDC, SVM, RLDA; secondly, it will use true error estimator with purpose to identify the most accurate classifier, which is the main goal of the project. Moreover, it will discuss and plot the results of error estimator by using different data sample sizes and dimensions. Finally, conclusion part will be provided

Table of Contents:

.

1. Introduction

In the modern era, when reading of a huge dataset is used to be something usual, it is important to find the most effective and accurate way to extract information from different types of data. The usage of extracted and analyzed information can be used in various spheres, such as at astronomy, economy and even at the crime combat (Cukier 2010). However, the problem with existing techniques is a limitation of variables, while large size of samples can be used. Therefore, in the last decades, it was vital to re-evaluate “well-known” ways, which are able to retrieve information from opposite situation with limited amount of samples and large number of variable

In the case of our report, datasets contain gene expression profiling in cancer from a group of people, which were divided into two groups with a good (class 1) and bad prognosis (class 0). At the first, it is wanted to identify the prognosis for a future patient to which group is referred. For that purpose classifiers (LDA, DLDA, QDA, G13, SEDC, SVM, RLDA), which will be described in the next part of the report, were used. At the second, it is needed to determine which classifier is the most accurate. Therefore, the true error estimator was applied, and the dataset of 5 different dimensions (p=3, 5, 10, 40, 80) for 4 various datasets (ChenLiver, NatsoulisRats, ZhanMyeloma, YeohLeukemia) were used. The results of applied methods will be shown on the graphs, and compared with each other.

2. System and Methods

2.1. Classifiers

2.1.1. Linear discriminant analysis (LDA)

Linear Discriminant Analysis (LDA) is one of the most famous and powerful classification method, which was developed in 1936 (Sayad 2010). It is widely used in statistic, pattern recognition and machine learning, to identify a samples’ linear combination that describes or divides two or more classes of objects.

In our case, it is defined that the future patients’ gene expression values for the same genes is plotted in a vector x (Anderson 1958). And, if the function of x (let’s call it

WLDA(x)) is larger than 0 then the patient is classified as 0 (bad prognosis); and if the value is less than 0, the patient is classified as 1 (good prognosis).

$$W^{LDA}(\mathbf{x}) = \left(\mathbf{x} - \frac{\bar{\mathbf{x}}_0 + \bar{\mathbf{x}}_1}{2}\right)^T \mathbf{C}^{-1}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1),$$

where,

$\bar{\mathbf{x}}_0 = \frac{1}{n_0}\sum_{\mathbf{x}_l \in S_0} \mathbf{x}_l$ and $\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{\mathbf{x}_l \in S_1} \mathbf{x}_l$ are the sample means of class 0 and 1. Also,

$$\mathbf{C} = \frac{(n_0 - 1)\mathbf{C}_0 + (n_1 - 1)\mathbf{C}_1}{n_0 + n_1 - 2},$$

where

$$\mathbf{C}_i = \frac{1}{n_i - 1}\sum_{\mathbf{x}_l \in S_i} (\mathbf{x}_l - \bar{\mathbf{x}}_i)(\mathbf{x}_l - \bar{\mathbf{x}}_i)^T.$$

Therefore, LDA classifier is determined by function $W_{LDA}(x)$ and is defined as:

$\psi_{LDA(X)} = \begin{cases} 1, if\ W_{LDA(x)} \leq c \\ 0, if\ W_{LDA(x)} > c \end{cases}$, where $c = \log\frac{(1-\alpha o)}{\alpha o}$ and α_o is probability to classify the sample to class 0.

2.1.2. Diagonal Linear Discriminant Analysis (DLDA)

Diagonal Linear Discriminant Analysis (DLDA) is more advanced type of LDA. What is more, DLDA is more effectively than other algorithms of LDA. In DLDA, D matrix has identical diagonal elements with pooled sample covariance matrix (C) in LDA, but the elements that are out of diagonal of D, equals to zero. The advantage of these qualifiers is that they are fast and even if there is a large number of samples.

$$W_{DLDF}(\mathbf{x}) = \left(\mathbf{x} - \frac{\bar{\mathbf{x}}_0 + \bar{\mathbf{x}}_1}{2}\right)^T \mathbf{D}^{-1}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1),$$

2.1.3. Qualitative Data Analysis (QDA)

In general, QDA is not so different from LDA. However, in QDA there is no supposition that the covariance matrix of each class must be the same.

Quadratic discriminant function:

$$W_{QDF}(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_0)^T(\alpha\times\mathbf{I}_p+\mathbf{C}_0)^{-1}(\mathbf{x}-\bar{\mathbf{x}}_0)+\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_1)^T(\alpha\times\mathbf{I}_p+\mathbf{C}_1)^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1)+\frac{1}{2}\log\left(\frac{|\alpha\times\mathbf{I}_p+\mathbf{C}_1|}{|\alpha\times\mathbf{I}_p+\mathbf{C}_0|}\right.$$

Perhaps it should also pointed out the fact that a quadratic function is very similar to the linear discriminant function except the covariance matrix, where an alpha constant is added to all diagonal elements. In our case the alpha is declared to 0.2 (generally, it should be small number greater than 0.1). The reason is that we cannot ignore the quadratic terms and will also include a second – order terms.

Classification rule:

G^(x)=arg max kδk(x)

One should note here that the classification rules are similar. In which maximum of quadratic discriminant function depends on class k.

Another good thing about QDA is that has a matrix to each class differently, therefore this classifier can be applied to the larger number of parameters. On the other hand, if there are considerable number of sample features that may lead to complication.

2.1.4. G13 Classifier

This classifier is very similar to the LDA and to obtain G13 only operation needed is multiply linear discriminant function by the expression $\frac{n_0+n_1-2-p}{n_0+n_1-2}$:

$$W_{G13}(x) = \left((x - \frac{\bar{x}_0+\bar{x}_1}{2}\right)^T C^{-1}\left(\bar{x}_0+\bar{x}_1\right)\left)\left(\frac{n_0+n_1-2-p}{n_0+n_1-2}\right)$$

Variables that used in the expression are described below:
n_0 and n_1 – sample points coming from populations 0 and 1 respectively;
p – dimension of column vectors x, $\bar{x}_0$ and $\bar{x}_1$

2.1.5. Sample Euclidean Classifier (SEDC Distance)

This modified version of linear discriminant analysis classifier (LDA) could be used, when the pooled sample covariance matrix is identity matrix *(C= I_p)* or it is necessary to ignore the information in it (Koolaard and Lawoko 1996, 2990).
The SEDC is defined by using the following function:

$$W_{SEDC}(x) = (z - \frac{\bar{z}_0+\bar{z}_1}{2})^T (\bar{z}_0 + \bar{z}_1)$$

On the discriminant above there are new vectors called z, $\bar{z}_0$ and $\bar{z}_1$ with dimension of m, where every element is average of consecutive l elements of p dimensional vectors x, $\bar{x}_0$ and $\bar{x}_1$ respectively. According to the statement above, z vectors have dimension m=p/l. For example, if p=60 and the number of considered consecutive variables is l=4, which is divisible to the p, then m=15. Therefore, z, $\bar{z}_0$ and $\bar{z}_1$ vectors can be illustrated as below:

$$z = \begin{bmatrix} \frac{x_1 + x_2 + x_3 + x_4}{4} \\ \frac{x_5 + x_6 + x_7 + x_8}{4} \\ \vdots \\ \frac{x_{p-3} + x_{p-2} + x_{p-1} + x_p}{4} \end{bmatrix}_{15}$$

$$z_0 = \begin{bmatrix} \frac{\bar{x}_{0.1}+\bar{x}_{0.2}+\bar{x}_{0.3}+\bar{x}_{0.4}}{4} \\ \frac{\bar{x}_{0.5}+\bar{x}_{0.6}+\bar{x}_{0.7}+\bar{x}_{0.8}}{4} \\ \vdots \\ \frac{\bar{x}_{0.p-3}+\bar{x}_{0.p-2}+\bar{x}_{0.p-1}+\bar{x}_{0.p}}{4} \end{bmatrix}_{15} \qquad z_1 = \begin{bmatrix} \frac{\bar{x}_{1.1}+\bar{x}_{1.2}+\bar{x}_{1.3}+\bar{x}_{1.4}}{4} \\ \frac{\bar{x}_{1.5}+\bar{x}_{1.6}+\bar{x}_{1.7}+\bar{x}_{1.8}}{4} \\ \vdots \\ \frac{\bar{x}_{1.p-3}+\bar{x}_{1.p-2}+\bar{x}_{1.p-1}+\bar{x}_{1.p}}{4} \end{bmatrix}_{15}$$

To evaluate the code, it was taken the value k=1, so m=p and vectors z, $\bar{z}_0$ and $\bar{z}_1$ are equivalent to mean value of vectors x, $\bar{x}_0$ and $\bar{x}_1$ respectively.

2.1.6. Support Vector Machine (SVM)

The support vector machine (SVM) is the learning method with the main aim is to separate a data into two classes, 1 and -1. Firstly, it finds a p-1-dimensional hyperplane, which divides the data of objects, represented as vectors in p-dimensional plane. Then, it finds a hyperplane with same conditions, but with larger margins i.e. with smaller distance between the hyperplane and nearest object. The process continues till it finds one with largest margin, which is called maximum-margin hyperplane. This hyperplane has equation: $w \cdot x - b = 0$, where w is orthogonal vector to the hyperplane, and the

parameter $\frac{b}{\|w\|}$ equals the distance to the origin, thus the value of w should be minimized to find maximum-margin hyperplane. So, for object $w \cdot x_i - b \geq 1$, then ci = 1, and if $w \cdot x_i - b \leq 1$, then ci = -1, or overall the equation will look like:

$c_i(w \cdot x_i - b) \geq 1$ (Boswell 2002).

2.1.7. Regularized Linear Discriminant Analysis (RLDA)

Regularized linear discriminant analysis (RLDA) is method similar to LDA with the same aim and almost the same algorithm. The only difference is an additional parameter γ, which is added to minimize the error. As it was written, the whole algorithm and threshold value are the same, the way of finding γ is: Firstly, the data is divided randomly into two sets, say A and B. In our code, they were two third and one third of the data, respectively. Then an arbitrary small value γ and maximum value γ_{max} are chosen to set a range, which was [0.001, 3^31]. With the exponential sequence gamma = 0.001*(3)^i, each gamma was trained in A data, and the best one with the least true error obtained was picked to use in B data (Friedman 1989).

$$W_{RLDF}(\mathbf{x}) = \left(\mathbf{x} - \frac{\bar{\mathbf{x}}_0 + \bar{\mathbf{x}}_1}{2}\right)^T (\gamma \times \mathbf{I}_p + \mathbf{C})^{-1}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)$$

3. Results and Discussion

This part of the report will discuss the accuracy of classifiers by considering the average of 500 numbers taken from the true error file in 5 different dimensions (p=3, 5, 10, 40, 80) and for 4 various datasets (ChenLiver, NatsoulisRats, ZhanMyeloma, YeohLeukemia). Results will be illustrated to each dataset separately.

3.1. ChenLiver

3.1.1. ChenLiver for p=3

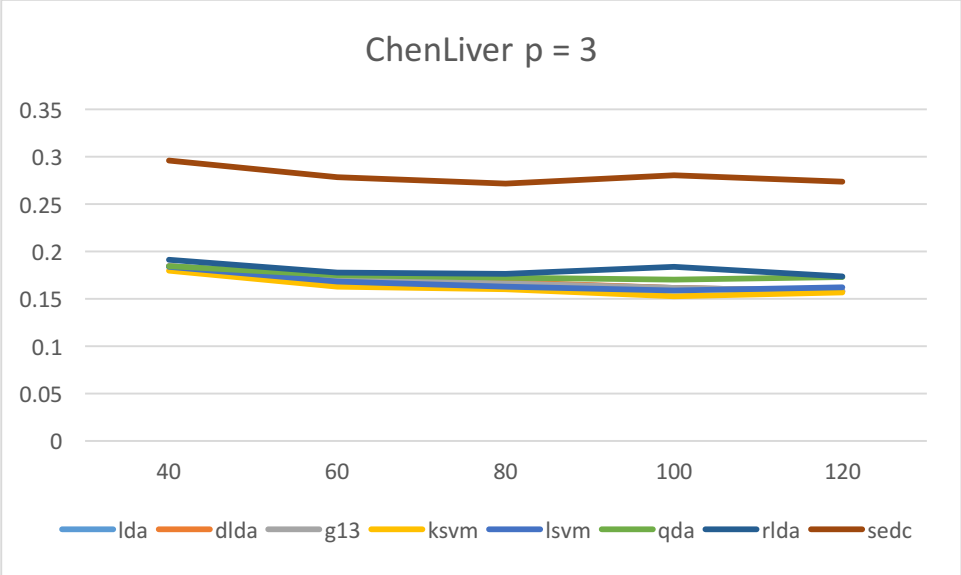

Figure 1. ChenLiver, p = 3

From the *Figure 1* it can be seen that KSVM has the smallest average for the true error value, when p=3. This has proven by considering the results for true error averages of each classifier:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0.168698	0.168605	0.168681	0.162573	0.167217	0.175004	0.180717	0.280036

3.1.2. ChenLiver for p=5

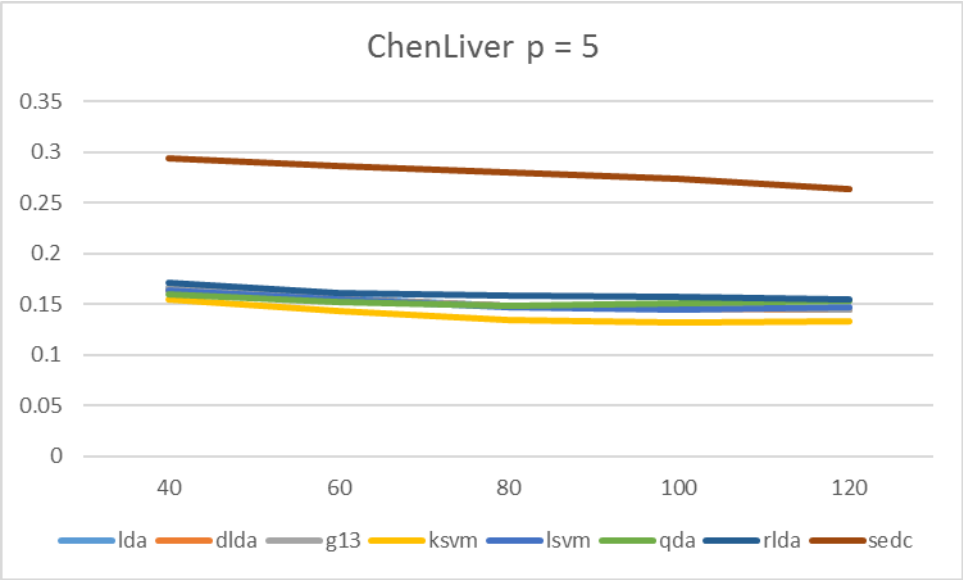

Figure 2. ChenLiver, p = 5

As shown on *Figure 2* KSVM has the highest accuracy for p=5 as well. This also justified by looking to the obtained data:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0.152246	0.151135	0.152122	0.139724	0.151157	0.152871	0.160605	0.279405

3.1.1. ChenLiver for p=10

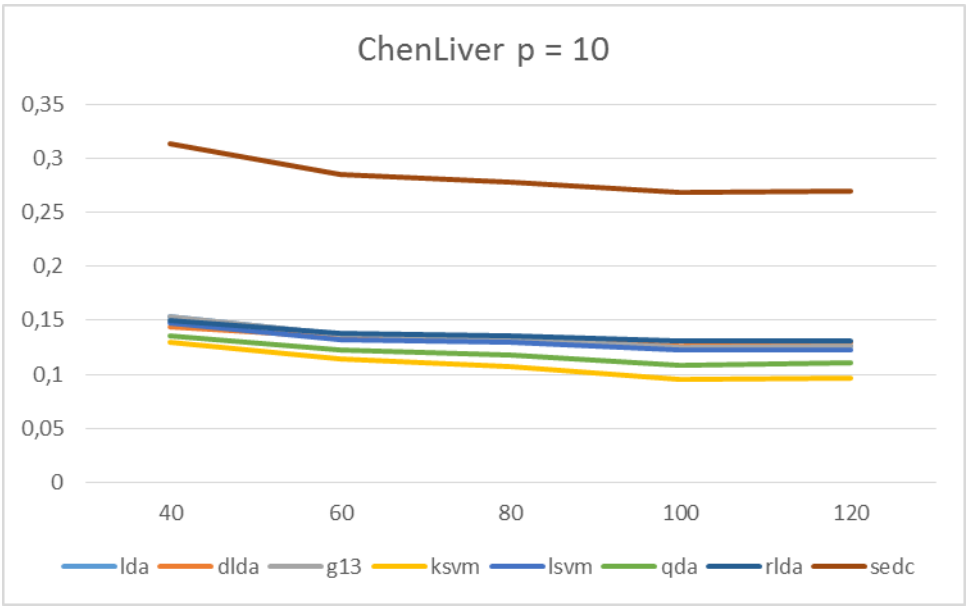

Figure 3. ChenLiver, p = 10

According to the graph constructed for the dimension 10, KSVM is more accurate compared to other classifiers. The data for averages is given below:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0.134997	0.133778	0.134524	0.108669	0.130945	0.119101	0.136912	0.282673

3.1.1. ChenLiver for p=40

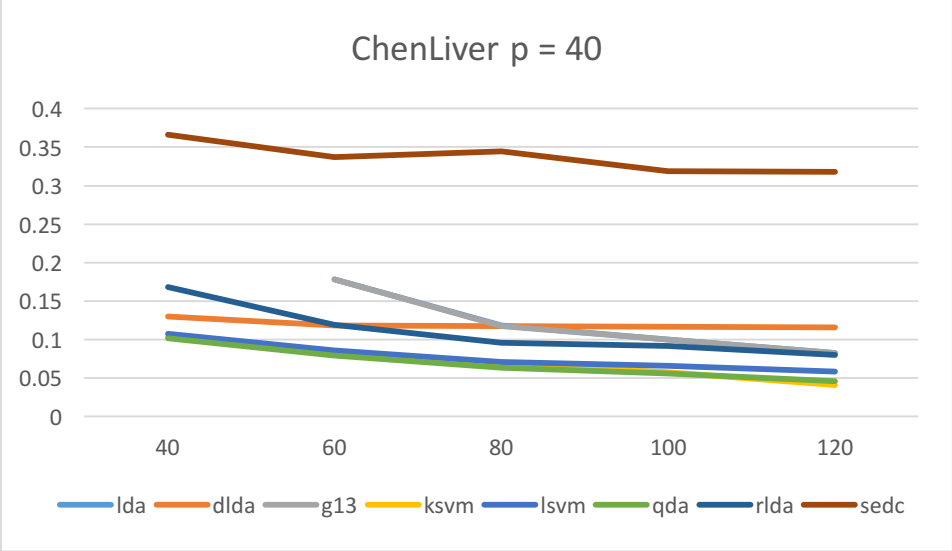

Figure 4. ChenLiver, p = 40

The graph for ChenLiver dataset when p=40 was plotted, and the curves for LDA and G13 started from n=60 (LDA and G13 work only when n>p). However, as there is no big difference in curves of two classifiers, namely QDA and KSVM, the table of results has been used to identify the most accurate one:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,119612	0,119536	0,11946	0,070806	0,077481	0,069193	0,110934	0,337016

From the table it can be determined that KSVM has the lowest value for true error, therefore has the highest value for accuracy.

3.1.2. ChenLiver for p=80

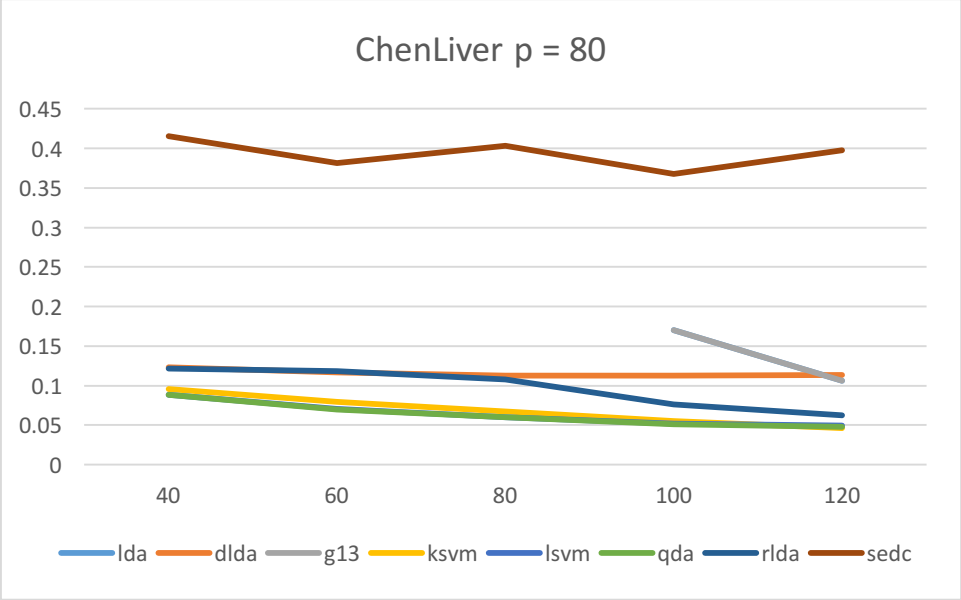

Figure 5. ChenLiver, p = 80

From the Figure 5 it is not hard to determine the classifier QDA has lowest value of true error. It was demonstrated on the table below as well:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,138328	0,1155638	0,137954	0,0688106	0,0639658	0,0632916	0,0973128	0,3932304

Total:

The KSVM had a lowest value of true error for 4 dimensions (when p=3,5,10,40), then it can be concluded that this classifier has a highest accuracy for ChenLiver dataset. Moreover, to ensure that this classifier is the most accurate for all dimensions, averages of true error values were found and compared:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0.142776	0.137724	0.142548	0.110117	0.118152	0.115892	0.137294	0.314537

$KSVM_{true}$ = 0.110117 (the lowest value of true error, the highest accuracy)

3.2. NatsoulisRats

3.2.1. NatsoulisRats for p=3

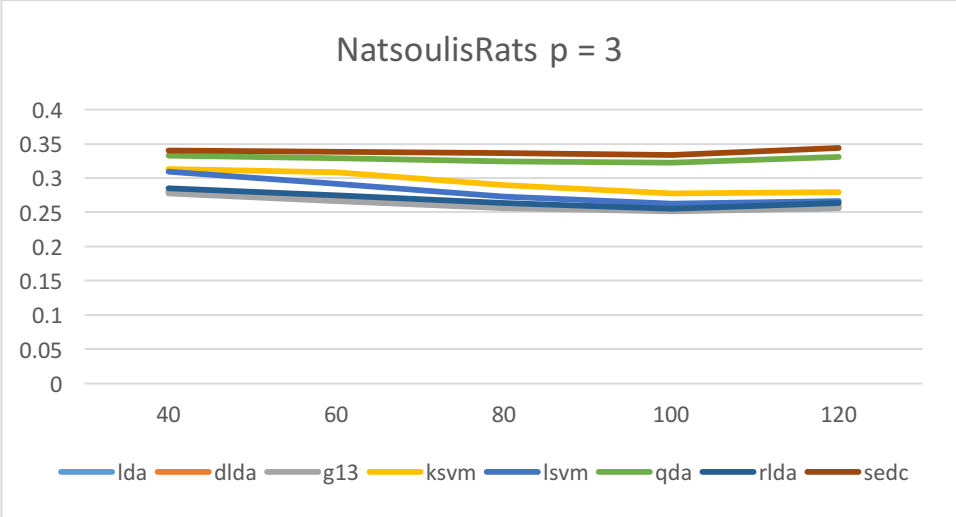

Figure 6. NatsoulisRats, p = 3

From the graph it can easily established that G13 classifier is more precise than others. The confirmation is given on the table below:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,262222	0,264968	0,261661	0,293871	0,280589	0,327822	0,26837	0,3387

3.2.2. NatsoulisRats for p=5

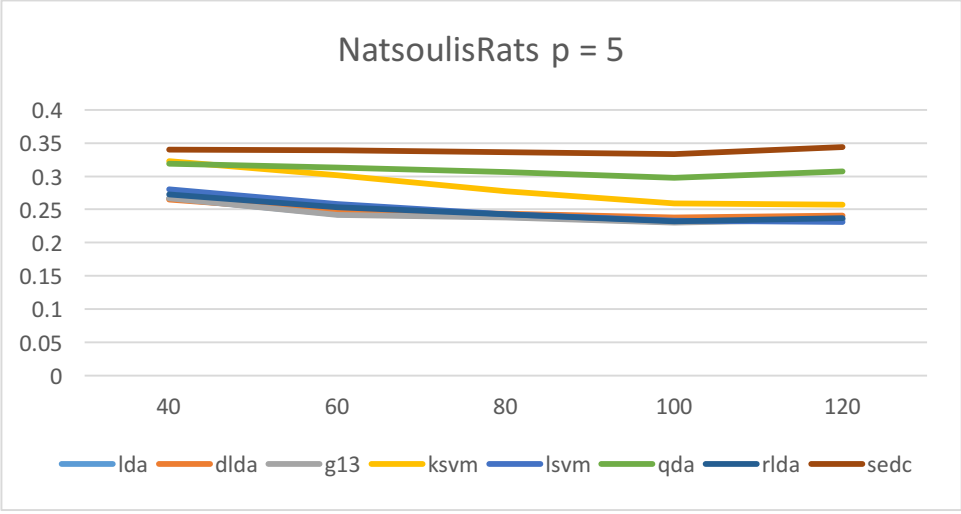

Figure 7. NatsoulisRats, p = 5

It is not clear which classifier has a higher accuracy, if use the graph only. Therefore, the table below was constructed and there was determined that G13 is the most precise classifier to this dimension:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,242713	0,246734	0,242318	0,283314	0,248836	0,308653	0,247663	0,3387

3.2.3. NatsoulisRats for p=10

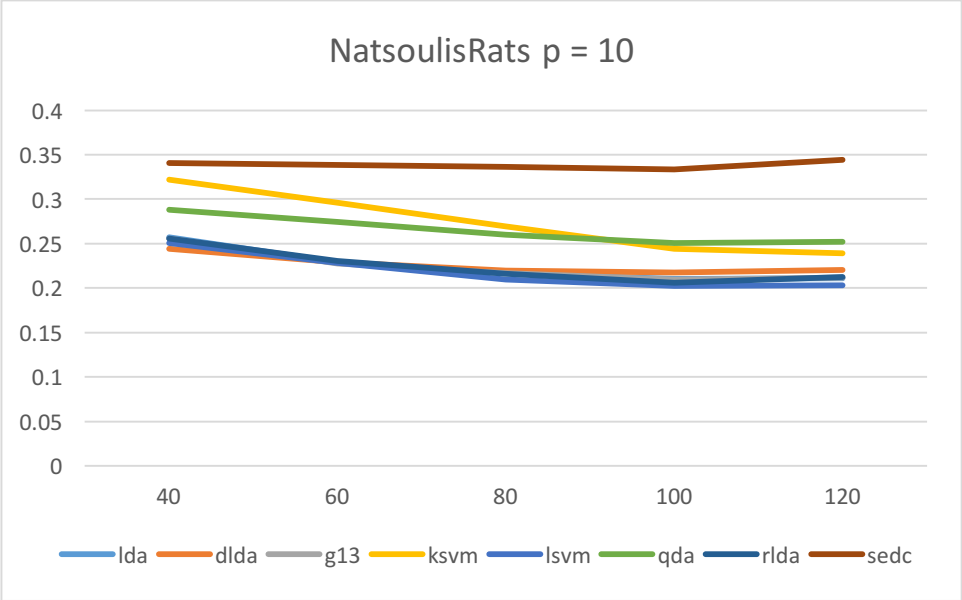

Figure 8. NatsoulisRats, p = 10

A classifier with highest accuracy from the graph: LSVM
Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,224947	0,226117	0,224239	0,274203	0,218842	0,265147	0,224089	0,3387

3.2.4. NatsoulisRats for p=40

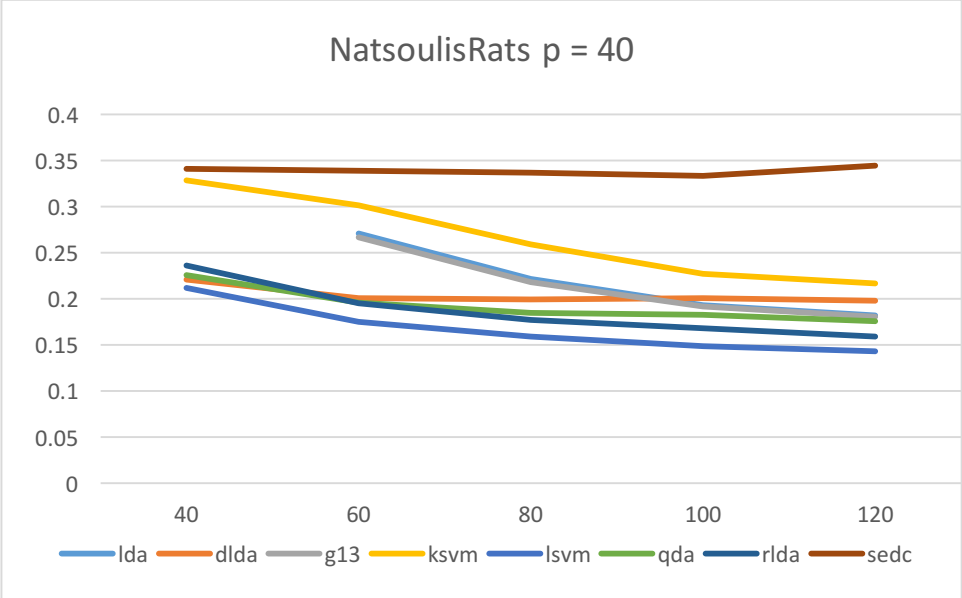

Figure 9. NatsoulisRats, p = 40

A classifier with highest accuracy from the graph: LSVM
Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,216704	0,203889	0,213962	0,266408	0,167483	0,192952	0,186986	0,3387

3.2.5. NatsoulisRats for p=80

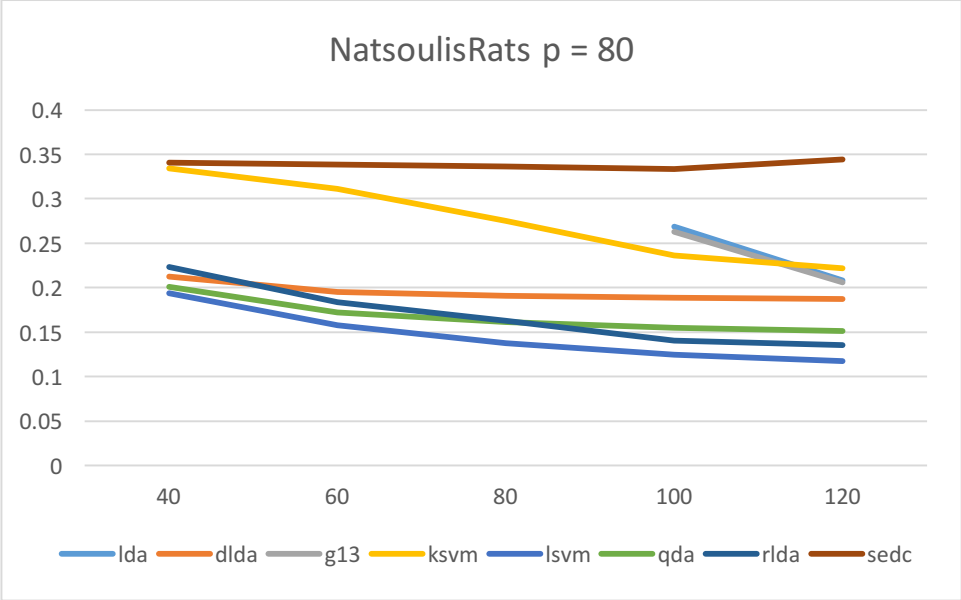

Figure 10. NatsoulisRats, p = 80

A classifier with highest accuracy from the graph: LSVM

Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,2382385	0,1949438	0,2342225	0,2756932	0,1462636	0,1682182	0,169225	0,3386996

Total:

According to the results taken from different dimensions, LSVM is the most precise classifier to the NatsoulisRats dataset.

Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,236965	0,227330	0,235281	0,278698	0,213463	0,255351	0,219267	0,33869

3.3. ZhanMyeloma

3.3.1. ZhanMyeloma for p=3

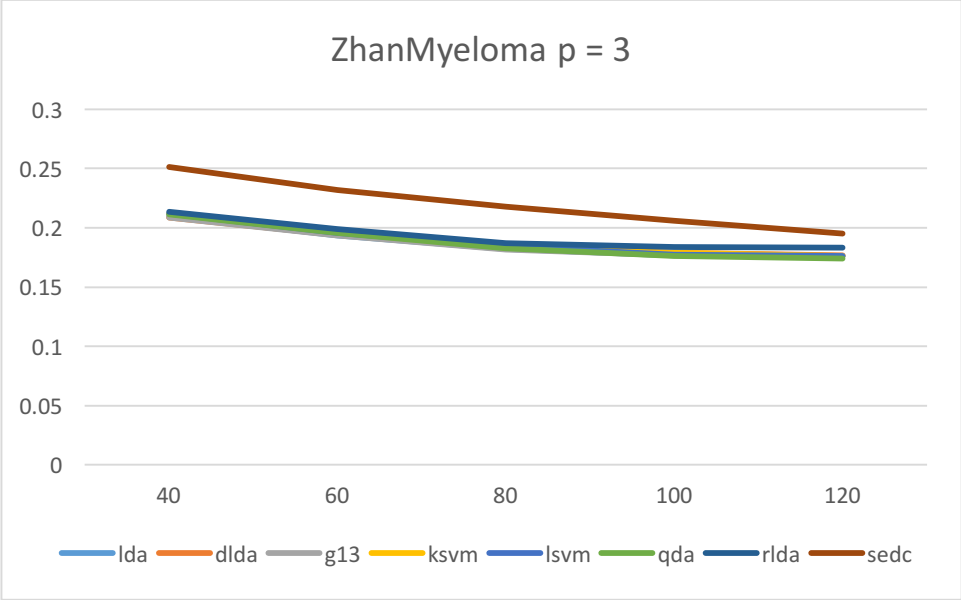

Figure 11. ZhanMyeloma, p = 3

A classifier with highest accuracy from the graph: not clear QDA or G13

Table of results to determine the most accurate classifier:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,187665	0,188323	0,18742	0,189186	0,189061	0,187946	0,193207	0,220313

3.3.2. ZhanMyeloma for p=5

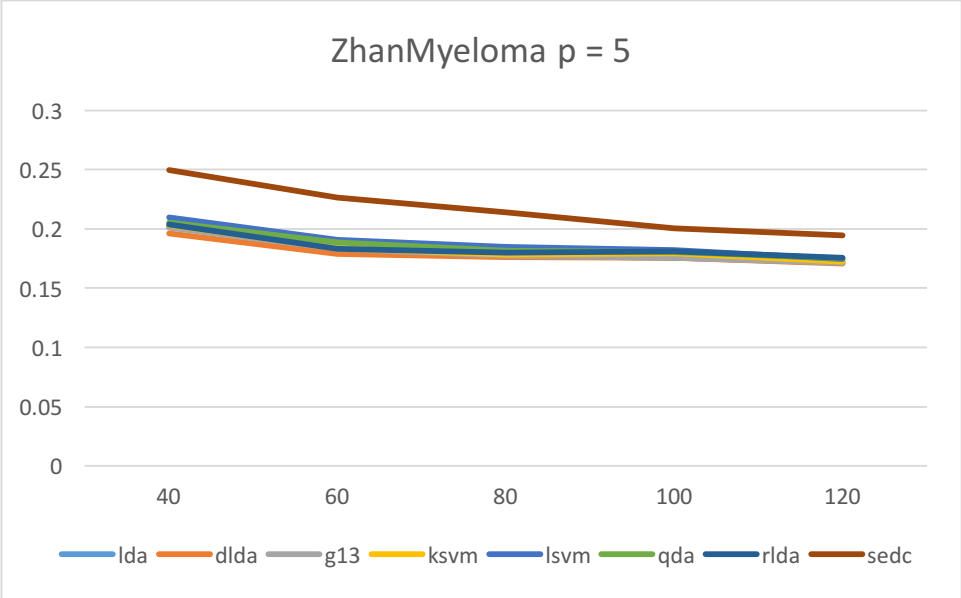

Figure 12. ZhanMyeloma, p = 5

A classifier with highest accuracy from the graph: DLDA

Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,182035	0,179517	0,181627	0,183326	0,188349	0,186204	0,184664	0,217052

3.3.3. ZhanMyeloma for p=10

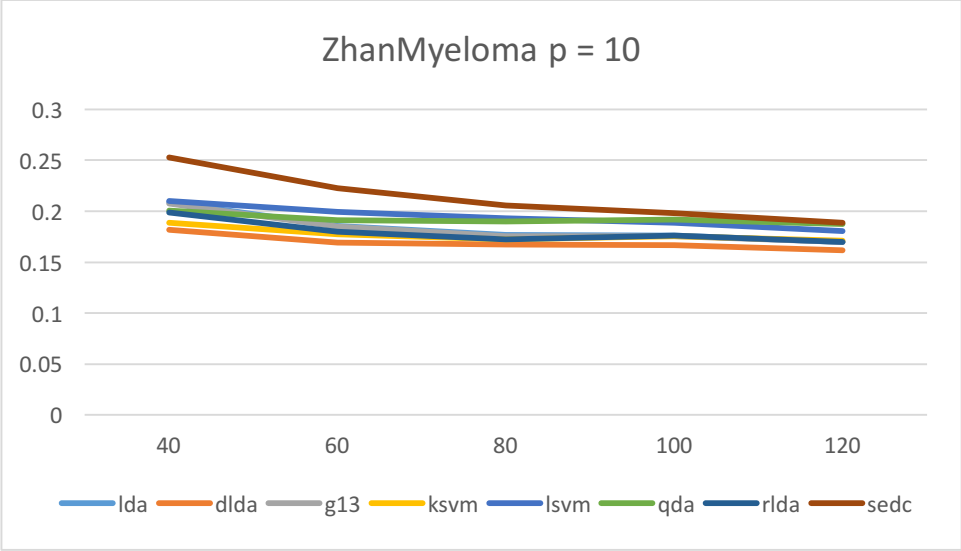

Figure 13. ZhanMyeloma, p = 10

A classifier with highest accuracy from the graph: DLDA

Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,183507	0,169405	0,182517	0,177066	0,194328	0,192086	0,179378	0,2135

3.3.4. ZhanMyeloma for p=40

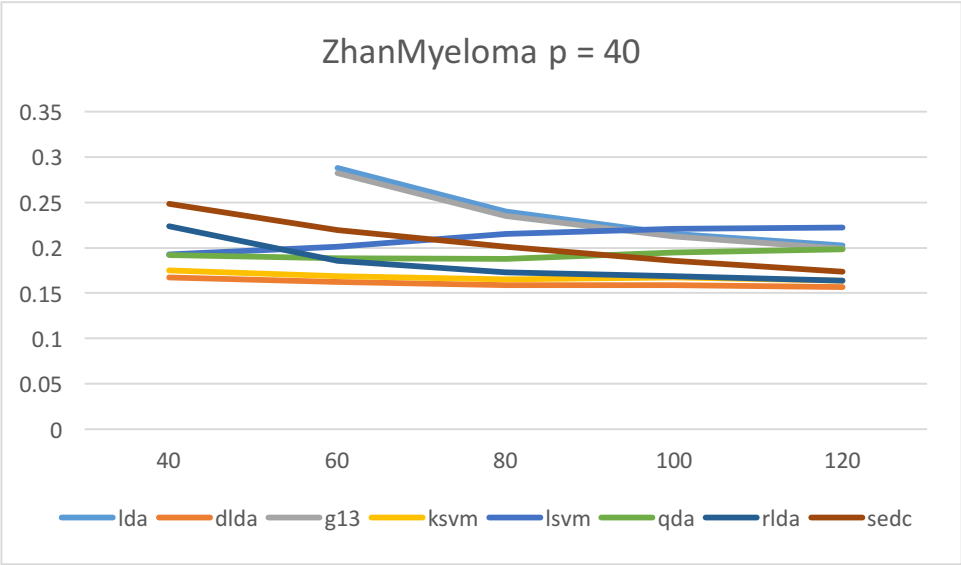

Figure 14. ZhanMyeloma, p = 40

A classifier with highest accuracy from the graph: DLDA

Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,236352	0,160685	0,23208	0,167896	0,210256	0,192215	0,183061	0,20562

3.3.5. ZhanMyeloma for p=80

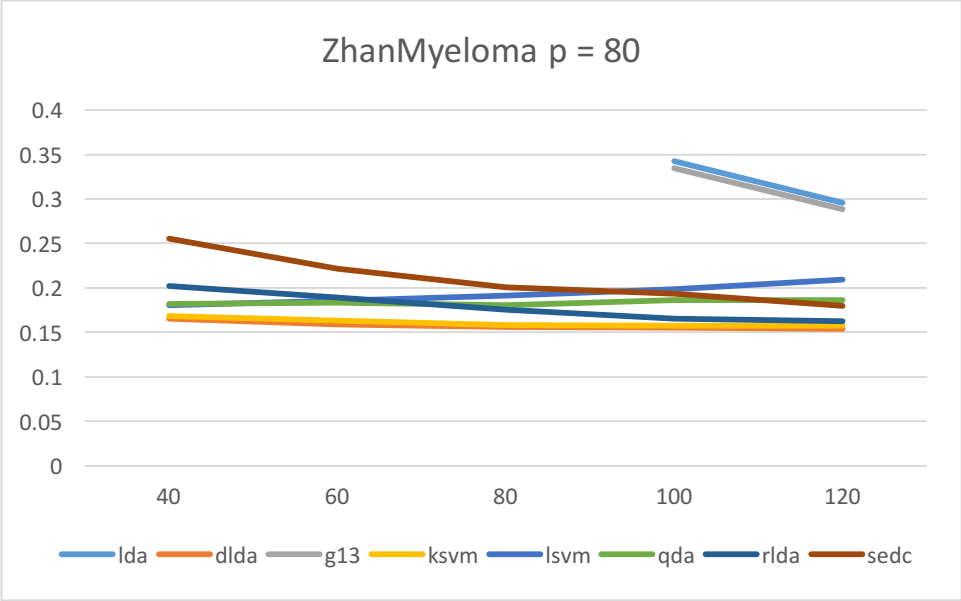

Figure 15. ZhanMyeloma, p = 80

A classifier with highest accuracy from the graph: DLDA

Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,318996	0,1578878	0,311611	0,1609234	0,1929702	0,1837684	0,1790292	0,21015

Total:

The Diagonal Linear Discriminant Analysis classifier has the lowest value for true error 4 dimensions, so it is the most accurate classification method for ZhanMyeloma dataset.

3.4. YeohLeukemia

3.4.1. YeohLeukemia for p=3

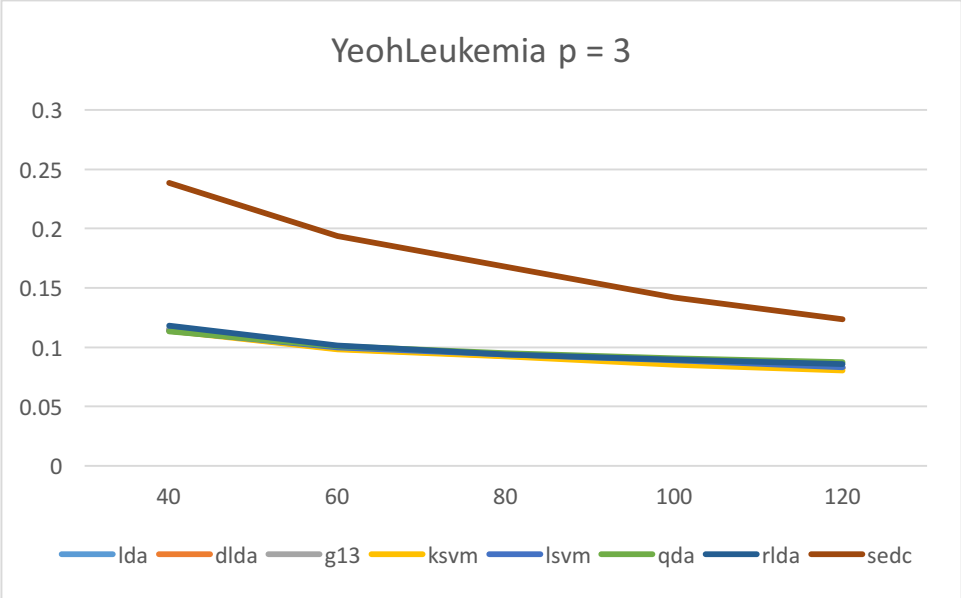

Figure 16. YeohLeukemia, p = 3

A classifier with highest accuracy from the graph: KSVM

Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,095666	0,095298	0,095654	0,093954	0,095928	0,097402	0,097706	0,173099

3.4.2. YeohLeukemia for p=5

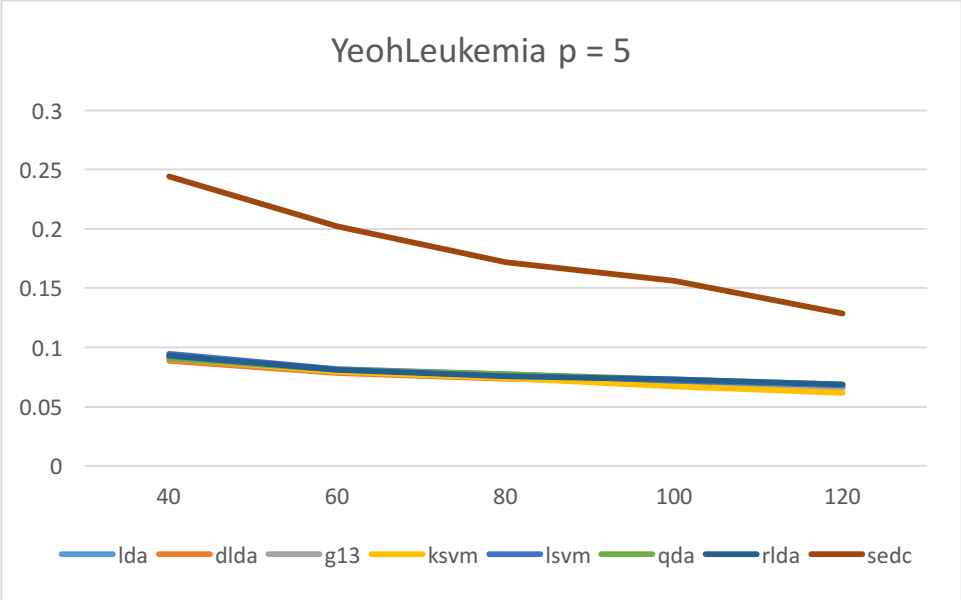

Figure 17. YeohLeukemia, p = 5

A classifier with highest accuracy from the graph: KSVM

Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,076927	0,0755	0,076977	0,074537	0,078695	0,078197	0,078136	0,180513

3.4.3. YeohLeukemia for p=10

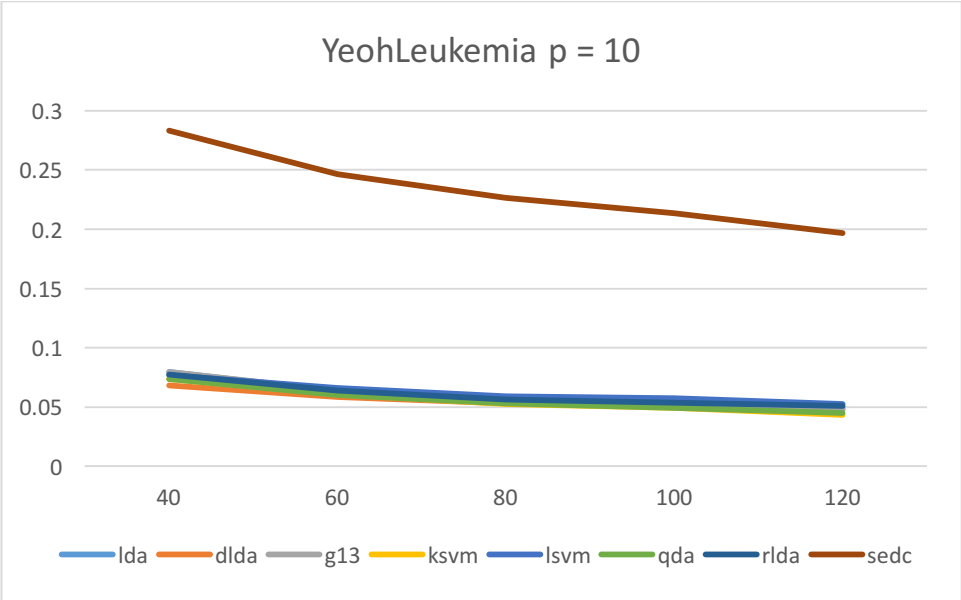

Figure 18. YeohLeukemia, p = 10

A classifier with highest accuracy from the graph: KSVM

Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,060059	0,05599	0,060173	0,055885	0,062359	0,056336	0,060397	0,233279

3.4.4. YeohLeukemia for p=40

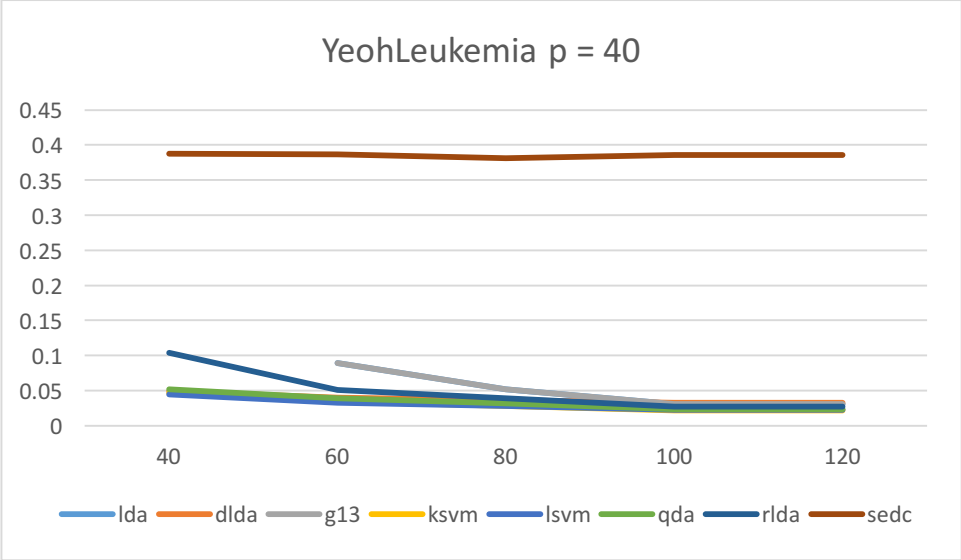

Figure 19. YeohLeukemia, p = 40

A classifier with highest accuracy from the graph: LSVM

Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,050684	0,038104	0,050722	0,030538	0,029816	0,033818	0,049392	0,385341

3.4.5. YeohLeukemia for p=80

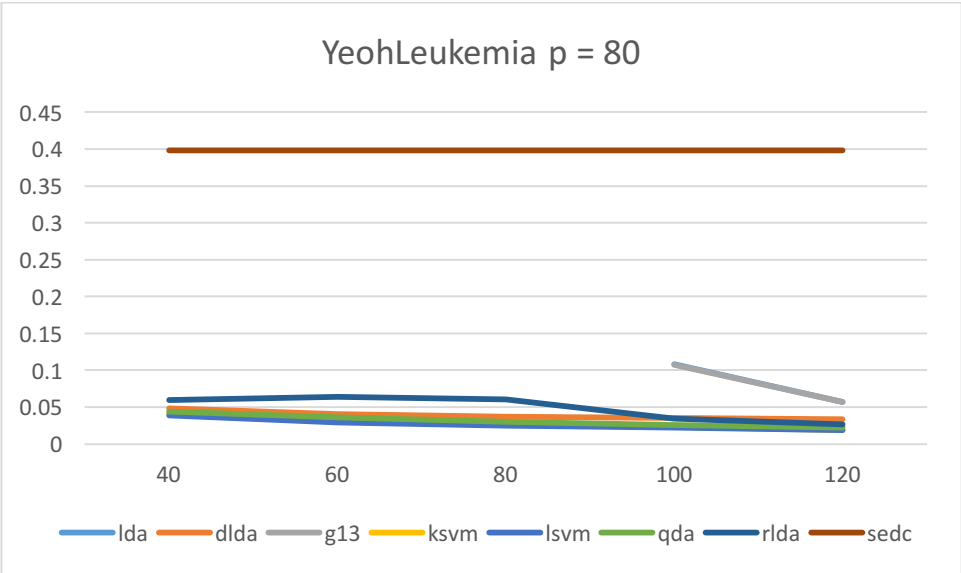

Figure 20. YeohLeukemia, p = 80

A classifier with highest accuracy from the graph: LSVM

Table of confirmation:

LDA	DLDA	G13	KSVM	LSVM	QDA	RLDA	SEDC
0,082504	0,0388296	0,0822415	0,0280624	0,02663	0,031515	0,048813	0,3983094

Overall:

For this dataset the linear and kernel SVM are the most accurate classifiers, therefore it can be concluded that SVM has the highest value of accuracy for YeohLeukemia.

4. Conclusion

The aim of this project was to identify the most accurate classifier. In order to achieve this goal, averages of classifier errors in various dimensions and data sizes were compared. By considering four datasets in five dimensions, twenty graphs of their true error were plotted. According to these graphs, it can be concluded that SVM classifier has the lowest deviation in most cases. Exception was found in results of ZhanMyeloma; DLDA classifier was the most accurate in this dataset. The diversity among the results could appear because of the different genes amount in each dataset, where ZhanMyeloma has the largest one (54,613). Finally, if it will be wanted to give prognosis for a future patient, it is more efficient to apply SVM, in order to reach higher accuracy.

Reference list:

Anderson, Theodor. 1958. *Introduction to Multivariate Statistical Analysis.* New York: Wiley.

Boswell, Dustin. 2002. Introduction To Support Vector Machines. Accessed November 19, 2016. http://dustwell.com/PastWork/IntroToSVM.pdf.

Cukier, Kenneth. 2010. “Data, data everywhere.” *The Economist*. Accessed November 20, 2016. http://www.economist.com/node/15557443

Friedman, Jerome H. 1989. "Regularized Discriminant Analysis". Journal Of The American Statistical Association 84 (405): 32. Accessed November 20, 2016. doi:10.2307/2289860.

Koolaard, John and C.R.O.Lawoko. 1996. “The linear and euclidean discriminant functions: A comparison via asymptotic expansions and simulation study.”*Communication in Statistics- Theory and Methods* 25(12):2989-3011. doi: 10.1080/03610929608831882

Sayad, Saed. 2010. “Linear Discriminant Analysis.” Accessed November 19, 2016. http://www.saedsayad.com/lda.htm