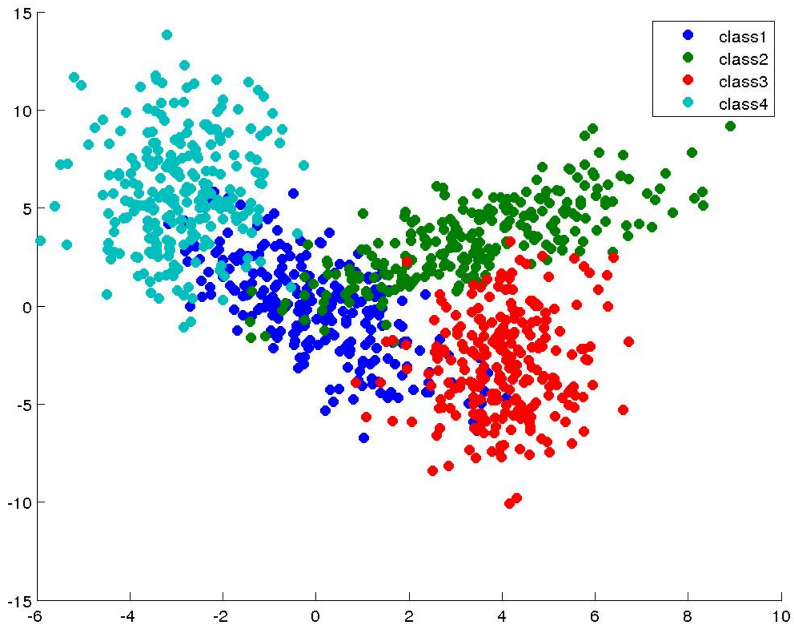
1 Introduction

2 Problem-1

3 Support Vector Data Description (SVDD) (one class SVM)

3.1 Dataset - 3 (2-dimensional overlapping data)

3.1.1 Data



The class-1 data (plotted in Blue color) from the dataset-3 is selected as 'normal data' to be learned and represented by one-class SVM.

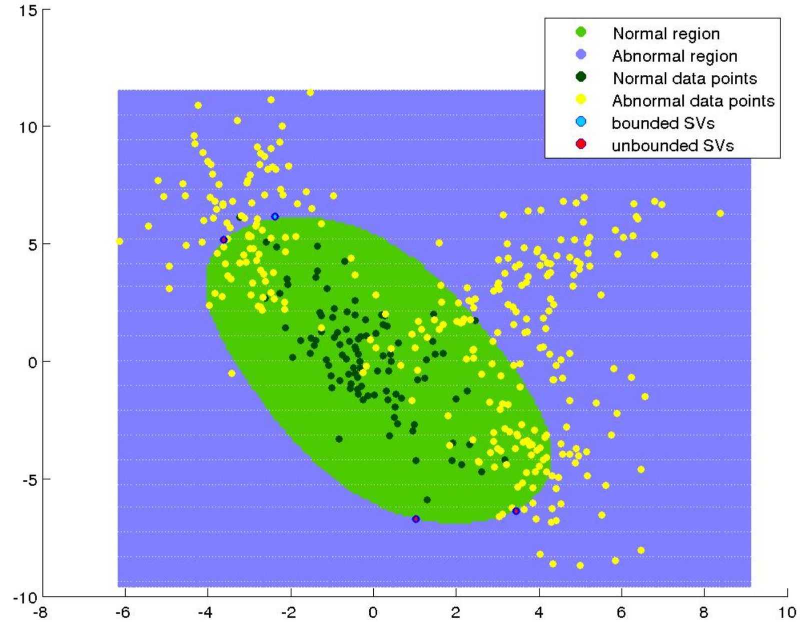
3.1.2 Procedure

- From the 2-dimensional overlapping data, one class (Class-1 with 250 data points) is selected as the 'normal' class used for training one class SVM.
- The data from other classes are uniformly sampled to form the validation (600 data points) and test set (400 data points), which are used to verify the efficiency of the trained model.
- The one-class SVM (with ν as hyperparameter) classifier is trained using different ν values and the best model is choosen based on the performance in validation set.

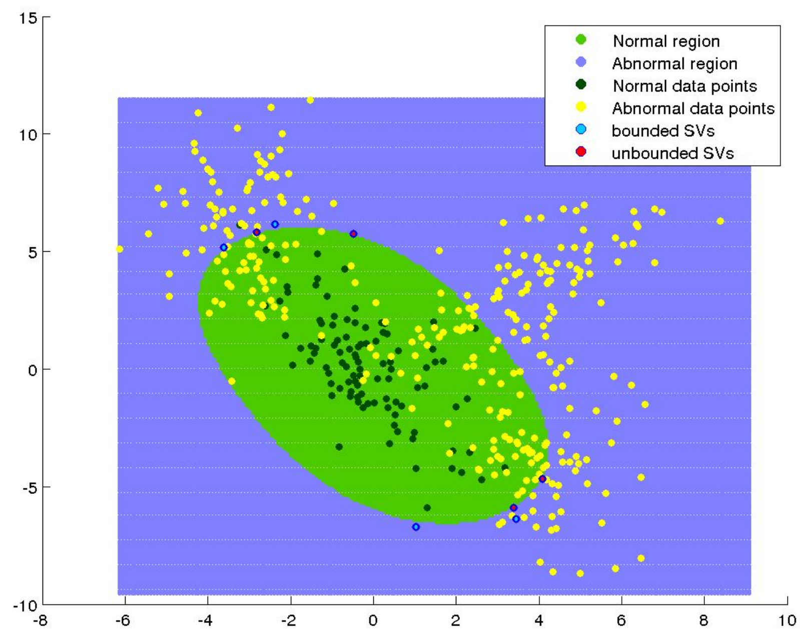
3.1.3 Decision regions

3.1.3.1 Illustration: Almost-Hard-minimal hypersphere SVDD

when $\nu = 0.01$, the SVDD model allows very less outliers. (i.e., Bounded support vectors = 1% of the training data).
Hence, the model trained by having ν as 0.01 has the disadvantage of classifying all other class data which are included inside the hypersphere, as the 'Normal' class. This behaviour will affect the performance of classifier, as shown in the figure 1.



(a) One-class SVM with RBF kernel (gamma = 0.01, $\nu = 0.01$, bounded SVs = 1, unbounded SVs = 3)



(b) One-class SVM with RBF kernel (gamma = 0.01, $\nu = 0.02$, bounded SVs = 4, unbounded SVs = 4)

Figure 1: illustration of Almost-hard-minimal hypersphere which gives poor performance

3.1.3.2 Confusion matrices & performance details

3.1.3.2.1 ν = 0.01, RBF kernel parameter (γ) = 0.01

Validation data

	Abnormal class (predicted)	Normal class (predicted)
Abnormal class (Target)	289	161
Normal class (Target)	4	146

	Formula	Value
True positive rate w.r.t., Normal class (recall)	$\frac{TP}{TP+FN}$	0.973
False positive rate w.r.t., Normal class (fall-out)	$\frac{FP}{TN+FP}$	0.357
F1-score	$\frac{2TP}{2TP+FN+FP}$	0.638
Accuracy		68.75%

Test data

	Abnormal class (predicted)	Normal class (predicted)
Abnormal class (Target)	176	124
Normal class (Target)	1	99

	Formula	Value
True positive rate w.r.t., Normal class (recall)	$\frac{TP}{TP+FN}$	0.999
False positive rate w.r.t., Normal class (fall-out)	$\frac{FP}{TN+FP}$	0.413
F1-score	$\frac{2TP}{2TP+FN+FP}$	0.613
Accuracy		68.75%

3.1.3.2.2 ν = 0.02, RBF kernel parameter (γ) = 0.01

Validation data

	Abnormal class (predicted)	Normal class (predicted)
Abnormal class (Target)	285	163
Normal class (Target)	5	145

	Formula	Value
True positive rate w.r.t., Normal class (recall)	$\frac{TP}{TP+FN}$	0.966
False positive rate w.r.t., Normal class (fall-out)	$\frac{FP}{TN+FP}$	0.362

F1-score	$\frac{2TP}{2TP+FN+FP}$	0.632
Accuracy		68.5%

Test data

	Abnormal class (predicted)	Normal class (predicted)
Abnormal class (Target)	175	125
Normal class (Target)	1	99

	Formula	Value
True positive rate w.r.t., Normal class (recall)	$\frac{TP}{TP+FN}$	0.99
False positive rate w.r.t., Normal class (fall-out)	$\frac{FP}{TN+FP}$	0.417
F1-score	$\frac{2TP}{2TP+FN+FP}$	0.611
Accuracy		68.75%

3.1.3.3 Soft-minimal hypersphere for better performance

To get an optimal soft-minimal hypersphere, ν can be set as 0.1 or 0.2 to allow 10% or 20% outliers (Bounded support vectors) of the data to lie out of the minimal hypersphere. This will give higher performance, as it will reduce the misclassification rate, as shown in the figure 2.

3.1.3.3.1 $\nu = 0.25$, RBF kernel parameter $(\gamma) = 0.01$

Validation data

	Abnormal class (predicted)	Normal class (predicted)
Abnormal class (Target)	408	42
Normal class (Target)	31	119

	Formula	Value
True positive rate w.r.t., Normal class (recall)	$\frac{TP}{TP+FN}$	0.793
False positive rate w.r.t., Normal class (fall-out)	$\frac{FP}{TN+FP}$	0.0933
F1-score	$\frac{2TP}{2TP+FN+FP}$	0.765
Accuracy		87.83%

Test data

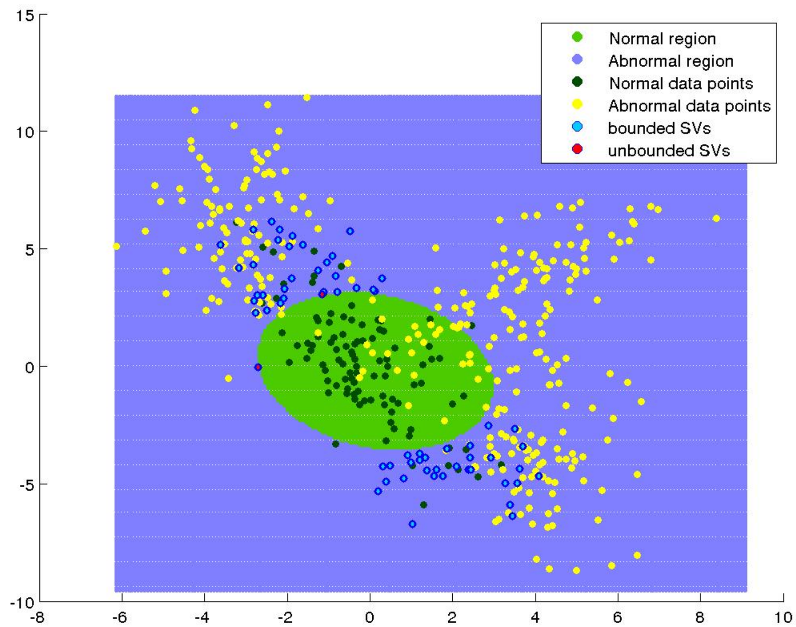


Figure 2: illustration of optimal soft-minimal hypersphere which gives good performance. One-class SVM with RBF kernel (gamma = 0.01, $\nu = 0.25$, bounded SVs = 61, unbounded SVs = 2)

	Abnormal class (predicted)	Normal class (predicted)
Abnormal class (Target)	269	31
Normal class (Target)	22	78

	Formula	Value
True positive rate w.r.t., Normal class (recall)	$\frac{TP}{TP+FN}$	0.78
False positive rate w.r.t., Normal class (fall-out)	$\frac{FP}{TN+FP}$	0.103
F1-score	$\frac{2TP}{2TP+FN+FP}$	0.746
Accuracy		86.75%

3.2 Dataset-4 : Breast benign

3.2.1 Procedure

- The given normal data (458 data points) is splitted into 70% as training set, 10% as validation set, the remaining 20% as test set.
- The given abnormal data (241 data points) is split into 50% as validation set and 50% as test set.
- One-class SVM (with RBF kernel) is trained only with Normal Data and the hyper-parameter(ν) is fixed using validation set.
- The efficiency of chosen model is evaluated using test set and results are reported.

The best validation performance of 94.54% is obtained, when RBF kernel parameter (γ) = 0.001 and $\nu = 0.05$, with Bounded SVs = 13, unbounded SVs = 4.

3.3 Confusion matrices and performance details

3.3.1 Validation data

	Abnormal class (predicted)	Normal class (predicted)
Abnormal class (Target)	116	4
Normal class (Target)	5	40

	Formula	Value
True positive rate w.r.t., Normal class (recall)	$\frac{TP}{TP+FN}$	0.889
False positive rate w.r.t., Normal class (fall-out)	$\frac{FP}{TN+FP}$	0.0334
F1-score	$\frac{2TP}{2TP+FN+FP}$	0.898
Accuracy		94.54%

3.3.2 Test data

	Abnormal class (predicted)	Normal class (predicted)
Abnormal class (Target)	118	3
Normal class (Target)	4	89

	Formula	Value
True positive rate w.r.t., Normal class (recall)	$\frac{TP}{TP+FN}$	0.956
False positive rate w.r.t., Normal class (fall-out)	$\frac{FP}{TN+FP}$	0.025
F1-score	$\frac{2TP}{2TP+FN+FP}$	0.962
Accuracy		96.72%

4 Problem-3

5 Problem-4

6 Problem-5 : Structured data classification

6.1 Task

The chosen task is to perform "classification" on Graphs for activity against non-small cell lung cancer and ovarian cancer cell lines.

6.2 Dataset

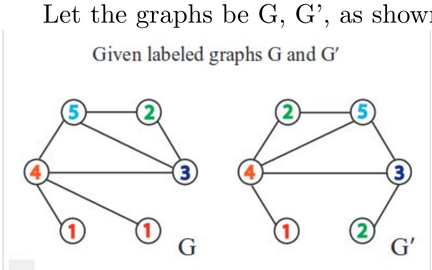
- The dataset NCI1 represent two balanced subsets of data sets of chemical compounds screened for activity against non-small cell lung cancer and ovarian cancer cell lines respectively (Wale and Karypis (2006) and `http://pubchem.ncbi.nlm.nih.gov`).
- The dataset contain 4110 instances of chemical compound activities represented as Graphs.
- The dataset is splitted into 70% as Training data, 10% as Validation data, 20% as Test data.

6.3 Kernel

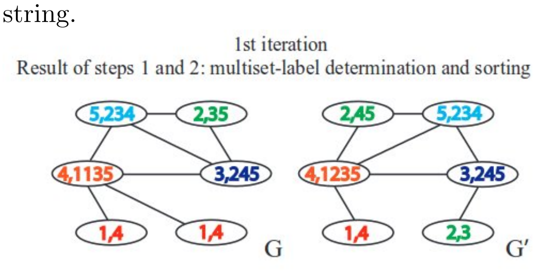
We use one of the graph kernel from Weisfeiler-Lehman (WL) Kernels family to define the similarity between two graphs. The kernel is called as "WL shortest path kernel". In essence, the shortest path kernel counts pairs of shortest paths with the same distance between identically labeled source and sink nodes on the original graphs.

All of the Weisfeiler-Lehman (WL) Kernels for given two graphs G, G' consists of basic 4 steps. as below:

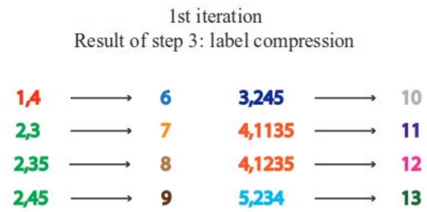
Let the graphs be G, G', as shown in the picture.



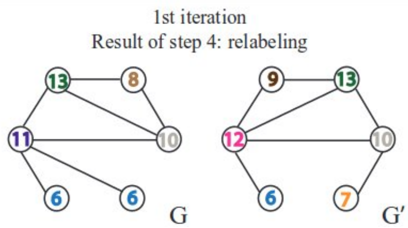
- **Multiset-label determination** : assign a multiset-label to each node in G and G' which consists of the multiset of neighborhood nodes.
- **Sorting each multiset** : Sort elements in the Multiset-label and concatenate them into a string.



- **Label compression** : Sort all of the strings for all $v \in G\&G'$ in ascending order. Map each distinct string into a new label using function $f : \sum^* \rightarrow \sum$, where $f(string(v)) = f(string(w))$, iff, $string(v) = string(w)$



- **Relabeling** : for all nodes in graphs G, G', set the label as $f(string(v))$



These four steps are repeated to get a final relabeled graph from which the Kernel's feature maps are derived.

6.3.0.1 Base kernel

we consider the base kernel k_{SP} of the form k_{SP}(G, G') = $\phi_{SP}(G)^T \phi_{SP}(G')$, where $\phi_{SP}(G)^T$ is a vector whose components are number of occurences of triplets of the form $< a, b, p >$ in G, where a,b $\in \sum$ are ordered end-point labels of shortest path and p $\in N_0$ is the shortest path length. The base kernel at each iteration i, can be combined together to give out the final kernel.

$k^h_{WLshortestpath} = k_{SP}(G_0, G'_0) + k_{SP}(G_1, G'_1) + \cdots + k_{SP}(G_h, G'_h)$

6.4 Confusion matrices and performance details

For the current dataset (NCI1), the iteration count is set as 3. The base kernel matrices for all the three iterations are retrieved and added to get the actual kernel $k^h_{WLshortestpath}$.

The best performance is achieved with $\nu = 0.3$. In the best model, the number of bounded SVs = 422, number of unbounded SVs = 1255.

6.4.1 Validation data

	non-small cell lung cancer (predicted)	ovarian cancer cell (predicted)
non-small cell lung cancer (Target)	181	24
ovarian cancer cell (Target)	30	175

	Formula	Value
True positive rate w.r.t., Normal class (recall)	$\frac{TP}{TP+FN}$	0.854
False positive rate w.r.t., Normal class (fall-out)	$\frac{FP}{TN+FP}$	0.218
F1-score	$\frac{2TP}{2TP+FN+FP}$	0.866
Accuracy		86.82%

6.4.2 Test data

	non-small cell lung cancer (predicted)	ovarian cancer cell (predicted)

non-small cell lung cancer (Target)	338	75
ovarian cancer cell (Target)	54	357

	Formula	Value
True positive rate w.r.t., Normal class (recall)	$\frac{TP}{TP+FN}$	0.868
False positive rate w.r.t., Normal class (fall-out)	$\frac{FP}{TN+FP}$	0.291
F1-score	$\frac{2TP}{2TP+FN+FP}$	0.846
Accuracy		84.34%