Semi-Supervised Learning
Dynamic Data

Nataliya Sokolovska

SPLEX, BIM, UPMC

Statistiques pour la classification et fouille de données en génomique

Outline

Semi-Supervised Learning

Kinetic Data

Stability Issues

Semi-Supervised Learning

Kinetic Data

Stability Issues

Semi-Supervised Learning

- Traditionally: Unsupervised and supervised learning
- Semi-supervised learning: halfway between supervised and unsupervised learning
- Semi-supervised learning with constraints: “these points have (or do not have) the same target”
- A problem related to SSL was introduced by V.Vapnik: transductive learning: do prediction for the test points only

A Brief History of Semi-Supervised Learning

- Self-learning: the earliest idea of SSL
  - Use repeatedly a supervised method.
  - It starts by training on the labeled data only; then label unlabeled data, etc.
- Transductive inference
  - Vapnik’s principle: When trying to solve some problem, one should not solve a more difficult problem as an intermediate step
  - No general decision rule is inferred
  - E.g. a combinatorial optimization on the labels of the test points in order to maximize the likelihood of their model
- Mixture of Gaussians
  - The likelihood of the model is maximized using the labeled and unlabeled data with the help of iterative algorithm such as Expectation-Maximization
  - Instead of mixture of Gaussians, use a mixture of multinomial distributions

A Brief History of Semi-Supervised Learning Cont’d

- Theoretical analysis
  - Learning rates exist for SSL of a mixture of two Gaussians: probability of error has an exponential convergence to the Bayes risk
- Text applications and natural language processing

When Can Semi-Supervised Learning Work?

In comparison with a supervised algorithm, can one hope to have a more accurate prediction by taking into account the unlabeled data?

- Prerequisite: the distribution of examples is relevant to the classification problem
- In a more mathematical formulation: the knowledge of $p(x)$ has to carry information that is useful in the inference of $p(y|x)$

When Can Semi-Supervised Learning Work?

The four assumptions:

- Smoothness assumption: If two points x_1 and x_2 are close, then so should be the corresponding y_1 and y_2
- Cluster assumption: If points are in the same cluster, they are likely to be of the same class
- Low density separation: The decision boundary should lie in a low-density region
- The (high-dimensional) data lie (roughly) on a low-dimensional manifold

Classes of Semi-Supervised Learning Algorithm

- Generative models
  - A generative model models $p(y, x)$, and any additional information on $p(x)$ is useful
  - It can be seen as classification with additional information on the marginal density
  - It can be seen as clustering with additional information
  - Advantage: Knowledge of the structure can be incorporated

Classes of Semi-Supervised Learning Algorithm

- Generative models
  - A generative model models $p(y, x)$, and any additional information on $p(x)$ is useful
  - It can be seen as classification with additional information on the marginal density
  - It can be seen as clustering with additional information
  - Advantage: Knowledge of the structure can be incorporated
- Low-density separation: an SVM
  - The most common approach – a maximum margin algorithm such as SVM
  - The method of maximizing the margin for unlabeled as well as labeled points is called the transduction SVM
  - The corresponding problem is non-convex, and thus difficult to optimize
- Low-density separation: entropy minimization
  - Encourage the class-conditional probability $p(y|x)$ to be close to 1 or to 0 at labeled and unlabeled points

Classes of Semi-Supervised Learning Algorithm

- Graph-based methods
  - Data are represented by the nodes of a graph, the edges of which are labeled with the pairwise distances of the incident nodes
  - Most graph methods use the graph Laplacian
  - Many graph methods penalize nonsmoothness along the edges
  - Intrinsically transductive and inductive algorithms
  - Information propagation on the graph

Classes of Semi-Supervised Learning Algorithm

- Graph-based methods
  - Data are represented by the nodes of a graph, the edges of which are labeled with the pairwise distances of the incident nodes
  - Most graph methods use the graph Laplacian
  - Many graph methods penalize nonsmoothness along the edges
  - Intrinsically transductive and inductive algorithms
  - Information propagation on the graph

- Change of Representation: two-step learning
  - Change representation: perform an unsupervised step on all data, and construct a new metric
  - Ignore the unlabeled data and perform supervised learning using the new data

Hypothesis and Notations

Notations:

- X_i observation
- Y_i label
- n the number of observation pairs
- $\pi(x,y)$ the joint probability
- $\eta(y|x)$ the conditional probability
- $q(x)$ the marginal probability of observations

The hypothesis:

- The marginal probability $q(x)$ is completely known
- $\mathcal{X}$ and $\mathcal{Y}$ are finite

Semi-Supervised Probabilistic Criterion

$\{X_i, Y_i\}_{i=1}^n$ are observations and their labels

Let $g(y|x;\theta)$ be the conditional probability function, parameterized by θ. Then the standard conditional maximum likelihood estimator is defined by

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i|X_i;\theta),$$

where $\ell(y|x;\theta) = -\log g(y|x;\theta)$ denotes the negated conditional log-likelihood function.

The asymptotically optimal semi-supervised estimator $\hat{\theta}_n^s$ is defined by

$$\hat{\theta}_n^s = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \frac{q(X_i)}{\sum_{j=1}^{n} \mathbb{1}\{X_j = X_i\}} \ell(Y_i|X_i;\theta),$$

where $q(x)$ is the marginal probability of observations.

Problem of the Covariate Shift

Covariate Shift

Let us learn an estimator from $(X_1,Y_1),\ldots,(X_n,Y_n)$, where the distribution of X_i is defined by $q_0(x)$. How to adapt the estimator if the test data X_i are distributed according to $q_1(x) \neq q_0(x)$?

- Si q_1 is known, the weights of the semi-supervised estimateur($q = q_1$) are asymototically identical to $\frac{1}{n}\frac{q_1}{q_0}(X_i)$ and the algorithm converges to

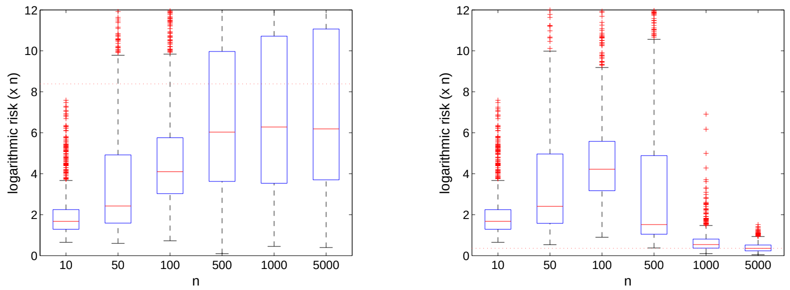
$$\theta_{1\star} = \arg\min_{\theta\in\Theta} \mathrm{E}_{\pi_1}[\ell(Y|X;\theta)] .$$

- The covariance matrix is smaller than the matrix of the estimator weighted by an importance ratio

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \frac{q_1}{q_0}(X_i)\ell(Y_i|X_i;\theta)$$

(which is supposed to know q_0).

Experiments with logistic regression



Boxplots of the scaled excess risk as a function of the number of observations in the presence of the covariate shift.
Left: Shimodaira criterion, $n\,(\mathrm{E}_\pi[\ell(Y|X;\hat{\theta}_n)] - \mathrm{E}_\pi[\ell(Y|X;\theta_\star)])$;
Right: semi-supervised estimator, $n\,(\mathrm{E}_\pi[\ell(Y|X;\hat{\theta}^s_n)] - \mathrm{E}_\pi[\ell(Y|X;\theta_\star)])$.

Applications to real problems

In the realistic applications (binary text classification), we can not assume that the true $q(x)$ is known.

We propose an approach based on clustering.

How to "estimate $q(x)$"? The set of unlabeled data is divided into k clusters, and in the expression of the weight

$$\frac{q(X_i)}{\sum_{j=1}^{n} \mathbb{1}\{X_j = X_i\}}$$

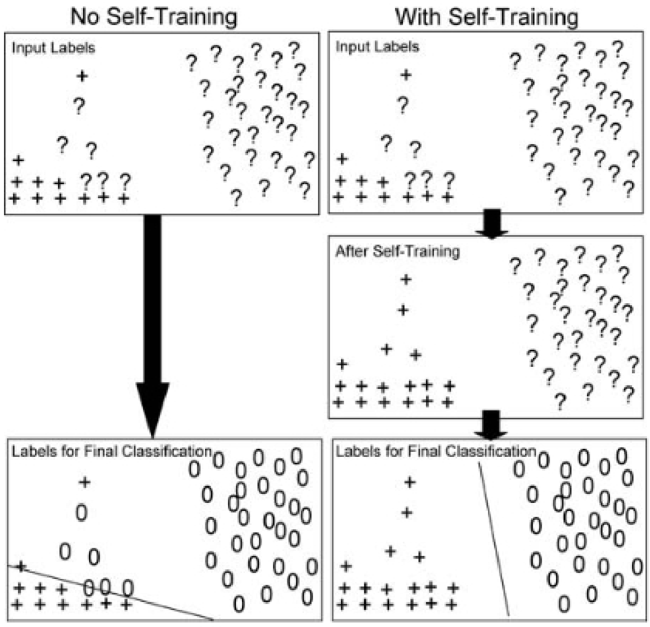
the numerator is replaced by the empirical frequency of the cluster which contains X_i; the denominator is replaced by the number of training points which are in the same cluster as X_i.

Example

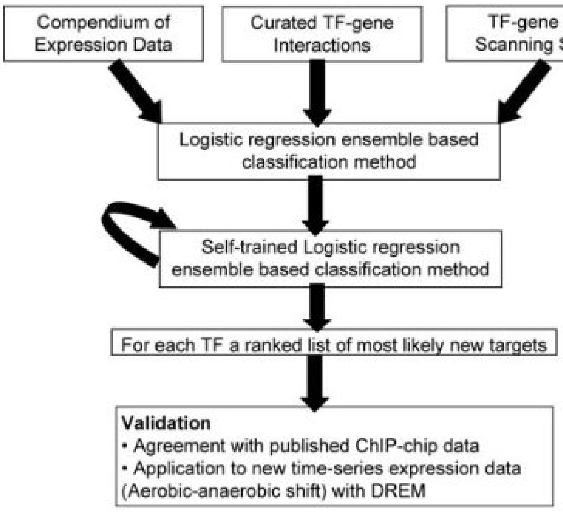
J. Ernst et al., A Semi-Supervised Method for Predicting Transcription Factor-Gene Interactions in Escherichia coli, PLOS, 2008 Problem:

- Try to combine gene expression and regulatory interactions to model transcriptional regulatory networks
- Using the available regulatory interactions to predict new interactions may lead to better coverage and more accurate models
- Use a database of verified transcriptional factor-gene interactions, DNA sequence binding motifs, and a compendium of gene expression data ⇒ predict new transcription factor-gene interactions

Method Overview



Method Overview



Compendium of Expression Data

Curated TF-gene Interactions

TF-gene Motif Scanning Scores

Logistic regression ensemble based classification method

Self-trained Logistic regression ensemble based classification method

For each TF a ranked list of most likely new targets

Validation

- Agreement with published ChIP-chip data
- Application to new time-series expression data (Aerobic-anaerobic shift) with DREM

Semi-Supervised Learning

Kinetic Data

Stability Issues

Kinetic Patterns

Ch. Baumgartner, A new data mining approach for profiling and categorizing kinetic patterns of metabolic biomarkers after myocardial injury, Bioinformatics, 2010

- Biomarkers have a substantial impact on the care of patients with cardiovascular disease
- They introduce a new evaluation model for prioritizing metabolic signatures in independent and dependent populations
- Perform ROC (receiver operating curve analysis) to estimate the power of the method

Kinetic Patterns

Ch. Baumgartner, A new data mining approach for profiling and categorizing kinetic patterns of metabolic biomarkers after myocardial injury, Bioinformatics, 2010

- Biomarkers have a substantial impact on the care of patients with cardiovascular disease
- They introduce a new evaluation model for prioritizing metabolic signatures in independent and dependent populations
- Perform ROC (receiver operating curve analysis) to estimate the power of the method

- 31 patients
- Data (blood samples) at 10 min, 60 min, 120 min, and 240 min (patients are not always the same) $\Rightarrow$ beyond the utility to study static phenotypes, one can choose a serial sampling design, and to look at kinetic relations
- Some data pre-processing is done (outliers detected, etc.)

Model for Paired samples

Paired Biomarker Identifier

$$pBI = \lambda * DA * \sqrt{\frac{|\Delta_{change}|}{|CV|}} * sign(\Delta_{change}),$$

- λ is a scaling factor
- DA is a discriminative measure
- CV is the coefficient of variation
- *sign* determines the direction of change

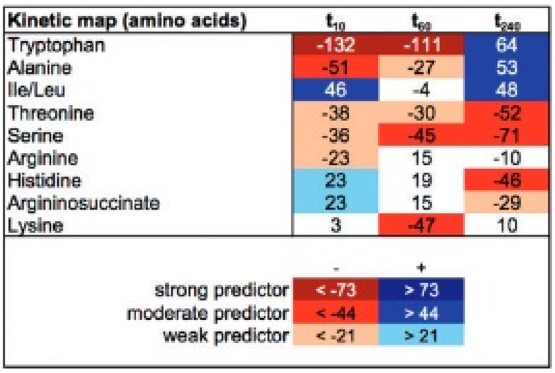
Model for Unpaired samples

Unpaired Biomarker Identifier: addresses the unpaired test problem, where we want to distinguish between two independent populations

$$uBI = \lambda * TP^2 * \sqrt{|\Delta_{change}|\frac{CV_{ref}}{CV}} * sign(\Delta_{change}),$$

- λ is a scaling factor
- TP^2 is a product of the true-positive rates of both classes (measure for discrimination)
- CV_{ref}/CV denotes changes in the variance of data across the two cohorts
- *sign* determines the direction of change

Kinetic Mapping



Kinetic map (amino acids)	t_{10}	t_{60}	t_{240}
Tryptophan	-132	-111	64
Alanine	-51	-27	53
Ile/Leu	46	-4	48
Threonine	-38	-30	-52
Serine	-36	-45	-71
Arginine	-23	15	-10
Histidine	23	19	-46
Argininosuccinate	23	15	-29
Lysine	3	-47	10

	-	+
strong predictor	< -73	> 73
moderate predictor	< -44	> 44
weak predictor	< -21	> 21

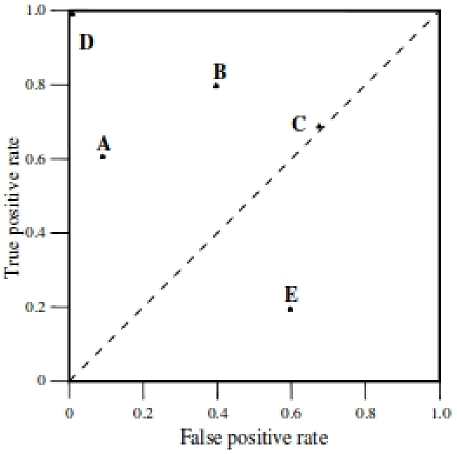
Kinetic map of amino acids on the data at 10, 60, and 240 min. after myocardial injury using the pBI scores. Red color indicates decreasing levels, and blue indicates increasing levels.

ROC analysis

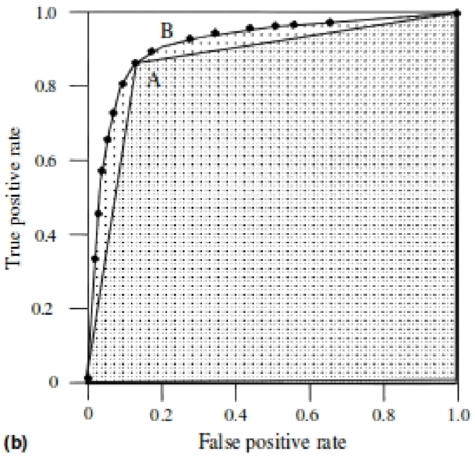
T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, 2006 A receiver operating characteristics (ROC) graph is a technique for visualizing, organizing and selecting classifiers based on their performance.

- ROC graphs are two-dimensional graphs in which *true positives* rate is plotted on the Y-axis and *false positives* rate is plotted on the X axis
- Probabilistic output classifiers
- Discrete output classifiers

ROC curves: example



ROC curves: example



AUC

- Area under a ROC curve
- The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance
- No realistic classifier should have an AUC less than 0.5

Semi-Supervised Learning

Kinetic Data

Stability Issues

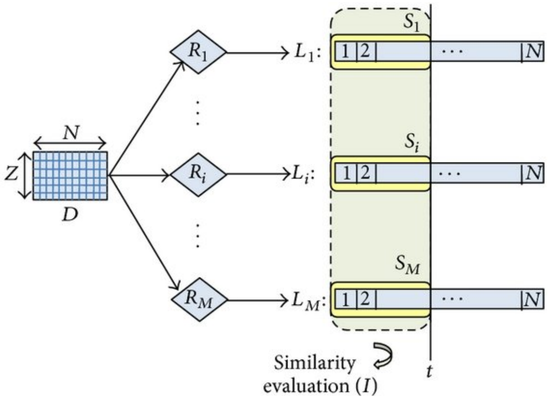
Biomarker Selection

N. Dessi et al., A comparative analysis of biomarker selection techniques, BioMed Research International, 2013

- Feature subset can be interpreted as a signature that captures significant knowledge for a given diagnostic task
- Aim is to compare, in a systematic way, the signatures produced by different rankers

Evaluate Similarity of Selected Gene Sets

Data set D with Z instances and N features (genes), a number M of rankers R_i ($i = 1, \ldots, M$) are applied to D. Each R_i produces a ranked list L_i where N features appear in descending order of relevance.



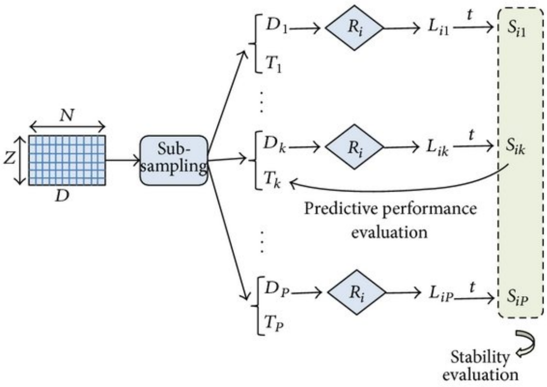
Evaluate Similarity of Selected Gene Sets

S_i are sets of genes. The similarity between two sets S_i and S_j can be expressed as the number of genes that are present in both sets, $|S_i \cap S_j|$, normalized to be in $[0; 1]$.

- Observation: the biological functions captured by different gene sets can be similar, despite a little degree of overlapping between these sets
- To exploit the gene sets in functional terms: gene annotations from the GO (Gene Ontology)
- GO provides a set of controlled vocabularies describing gene products based on their functions in the cell
- For each gene set S_i extract the list of molecular functions

Joint evaluation of stability and predictive performance

Extract from the original dataset D with Z instances and N features, a number P of reduced data D_k, $k = 1, \ldots P$, each containing $f * Z$ instances randomly drawn from D.



Joint evaluation of stability and predictive performance

To incorporate predictive performance evaluation in the above experimental protocol, build on each reduced data set a classification model.

Data sets

- Colon Tumor data set; containing 62 biological samples distinguished between tumor colon tissues (40 samples) and normal colon tissues (22 samples); each sample is described by the expression level of 2000 genes
- Leukemia dataset, containing 72 samples belonging to patients suffering from acute myeloid leukemia (25 samples) and acute lymphoblastic leukemia (47 samples); each sample is described by the expression level of 7129 genes.

Predictive performance

Methods tested: univariate techniques

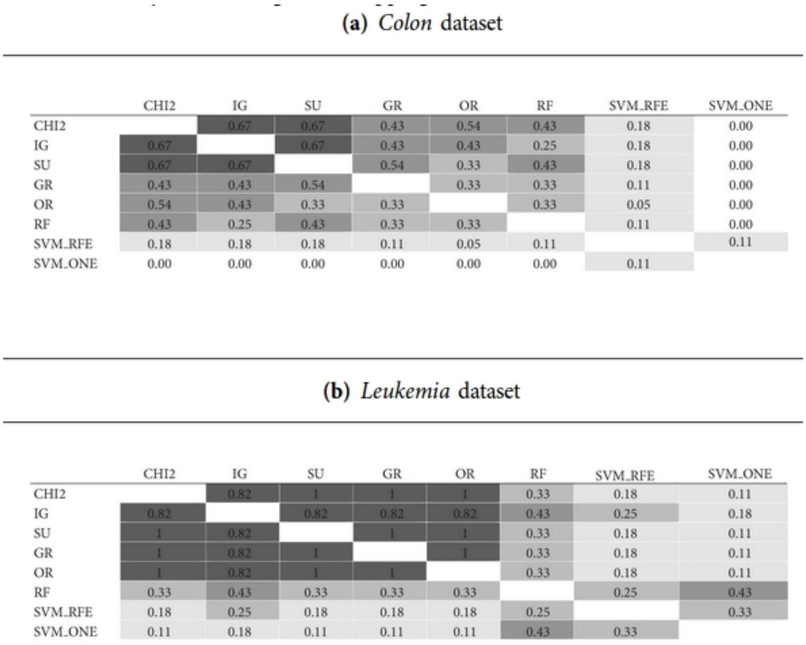
- chi Squared
- information gain
- symmetrical uncertainty
- gain ratio
- oneR

Predictive performance

Methods tested: multivariate techniques

- ReliefF
- SVM-embedded feature selection

Evaluate Similarity of Selected Gene Sets



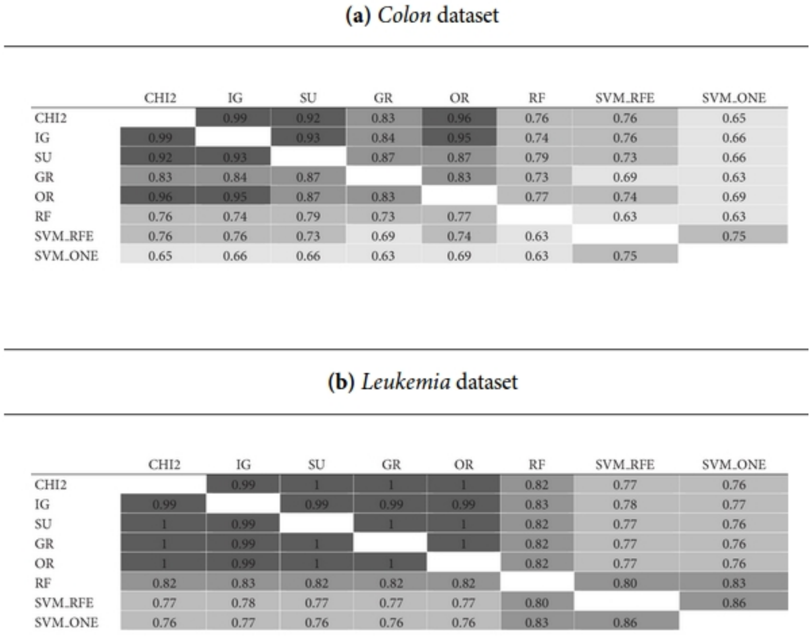
(a) *Colon* dataset

	CHI2	IG	SU	GR	OR	RF	SVM_RFE	SVM_ONE
CHI2		0.67	0.67	0.43	0.54	0.43	0.18	0.00
IG	0.67		0.67	0.43	0.43	0.25	0.18	0.00
SU	0.67	0.67		0.54	0.33	0.43	0.18	0.00
GR	0.43	0.43	0.54		0.33	0.33	0.11	0.00
OR	0.54	0.43	0.33	0.33		0.33	0.05	0.00
RF	0.43	0.25	0.43	0.33	0.33		0.11	0.00
SVM_RFE	0.18	0.18	0.18	0.11	0.05	0.11		0.11
SVM_ONE	0.00	0.00	0.00	0.00	0.00	0.00	0.11	

(b) *Leukemia* dataset

	CHI2	IG	SU	GR	OR	RF	SVM_RFE	SVM_ONE
CHI2		0.82	1	1	1	0.33	0.18	0.11
IG	0.82		0.82	0.82	0.82	0.43	0.25	0.18
SU	1	0.82		1	1	0.33	0.18	0.11
GR	1	0.82	1		1	0.33	0.18	0.11
OR	1	0.82	1	1		0.33	0.18	0.11
RF	0.33	0.43	0.33	0.33	0.33		0.25	0.43
SVM_RFE	0.18	0.25	0.18	0.18	0.18	0.25		0.33
SVM_ONE	0.11	0.18	0.11	0.11	0.11	0.43	0.33	

Evaluate Similarity of Selected Gene Sets



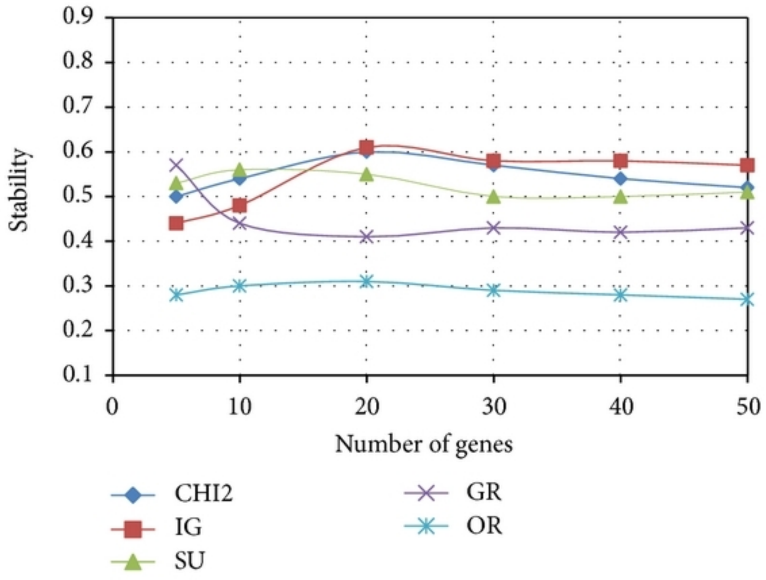
(a) *Colon* dataset

	CHI2	IG	SU	GR	OR	RF	SVM_RFE	SVM_ONE
CHI2		0.99	0.92	0.83	0.96	0.76	0.76	0.65
IG	0.99		0.93	0.84	0.95	0.74	0.76	0.66
SU	0.92	0.93		0.87	0.87	0.79	0.73	0.66
GR	0.83	0.84	0.87		0.83	0.73	0.69	0.63
OR	0.96	0.95	0.87	0.83		0.77	0.74	0.69
RF	0.76	0.74	0.79	0.73	0.77		0.63	0.63
SVM_RFE	0.76	0.76	0.73	0.69	0.74	0.63		0.75
SVM_ONE	0.65	0.66	0.66	0.63	0.69	0.63	0.75	

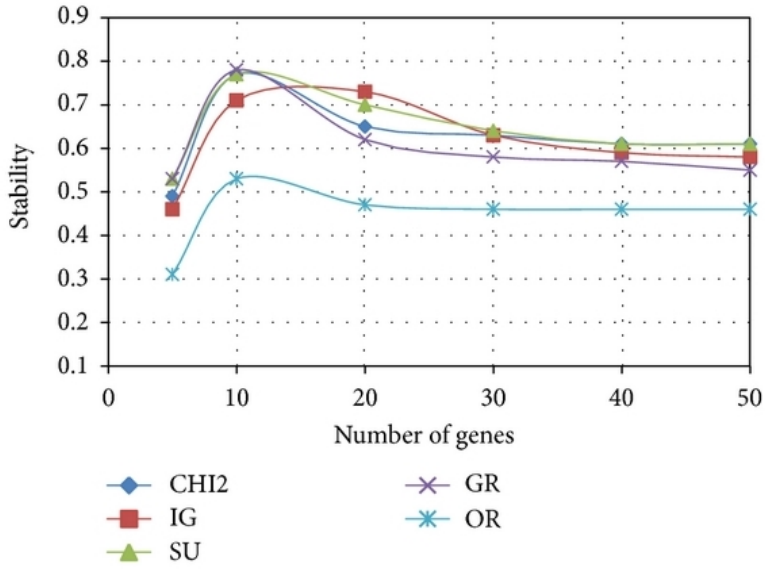
(b) *Leukemia* dataset

	CHI2	IG	SU	GR	OR	RF	SVM_RFE	SVM_ONE
CHI2		0.99	1	1	1	0.82	0.77	0.76
IG	0.99		0.99	0.99	0.99	0.83	0.78	0.77
SU	1	0.99		1	1	0.82	0.77	0.76
GR	1	0.99	1		1	0.82	0.77	0.76
OR	1	0.99	1	1		0.82	0.77	0.76
RF	0.82	0.83	0.82	0.82	0.82		0.80	0.83
SVM_RFE	0.77	0.78	0.77	0.77	0.77	0.80		0.86
SVM_ONE	0.76	0.77	0.76	0.76	0.76	0.83	0.86	

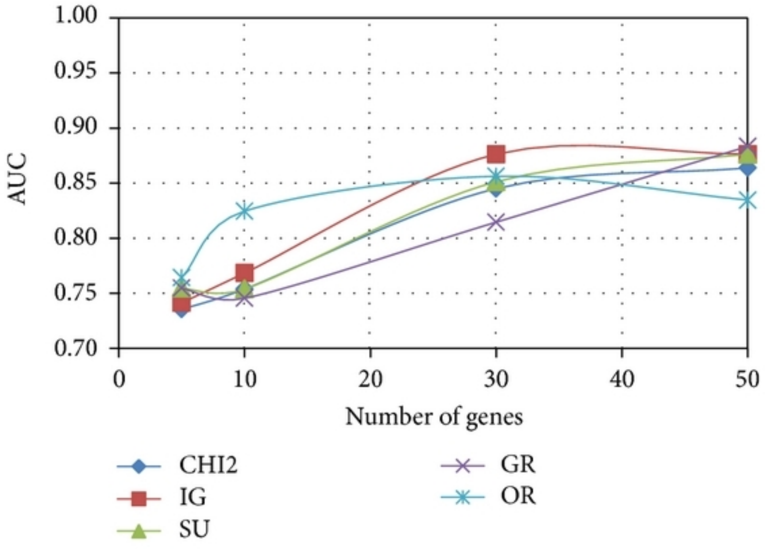
Evaluate Similarity of Selected Gene Sets



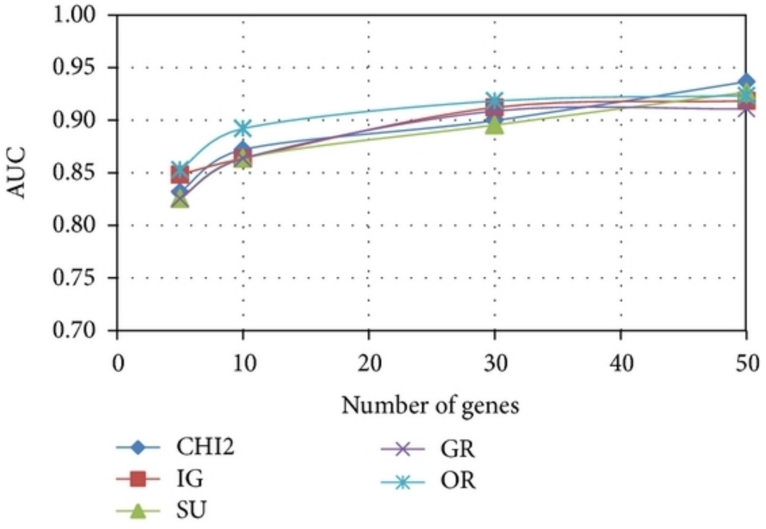
Evaluate Similarity of Selected Gene Sets



Similarity in Terms of Genes Overlapping



Functional Similarity


AUC
Number of genes
CHI2
IG
SU
GR
OR