O uso de conjuntos de dados de expressão gênica na pesquisa de seleção de atributos

Bruno Iochins Grisci

Feature Selection

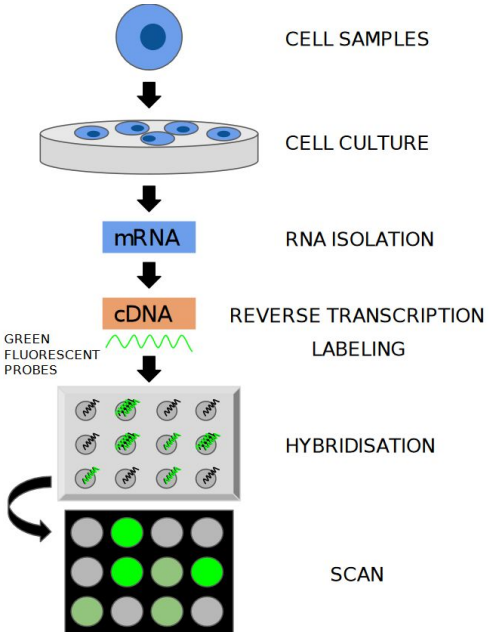
Generative AI Academy

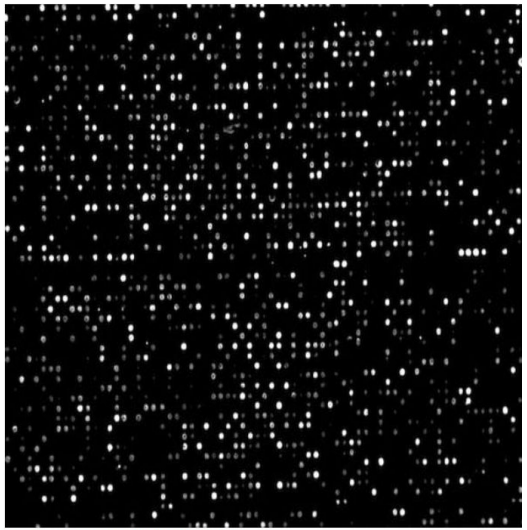
Part 2: Benchmark datasets

INSTITUTO DE INFORMÁTICA
UFRGS

Dados de expressão gênica

- A estrutura e função de diferentes tipos de células do mesmo organismo são diferentes, enquanto o genoma permanece o mesmo.
- Diferentes genes em cada tipo de célula são expressos.
- Pode ser usado para estudar a diferença na expressão entre tecidos do mesmo tipo sob diferentes condições.
- Novos biomarcadores podem ser potenciais alvos terapêuticos.



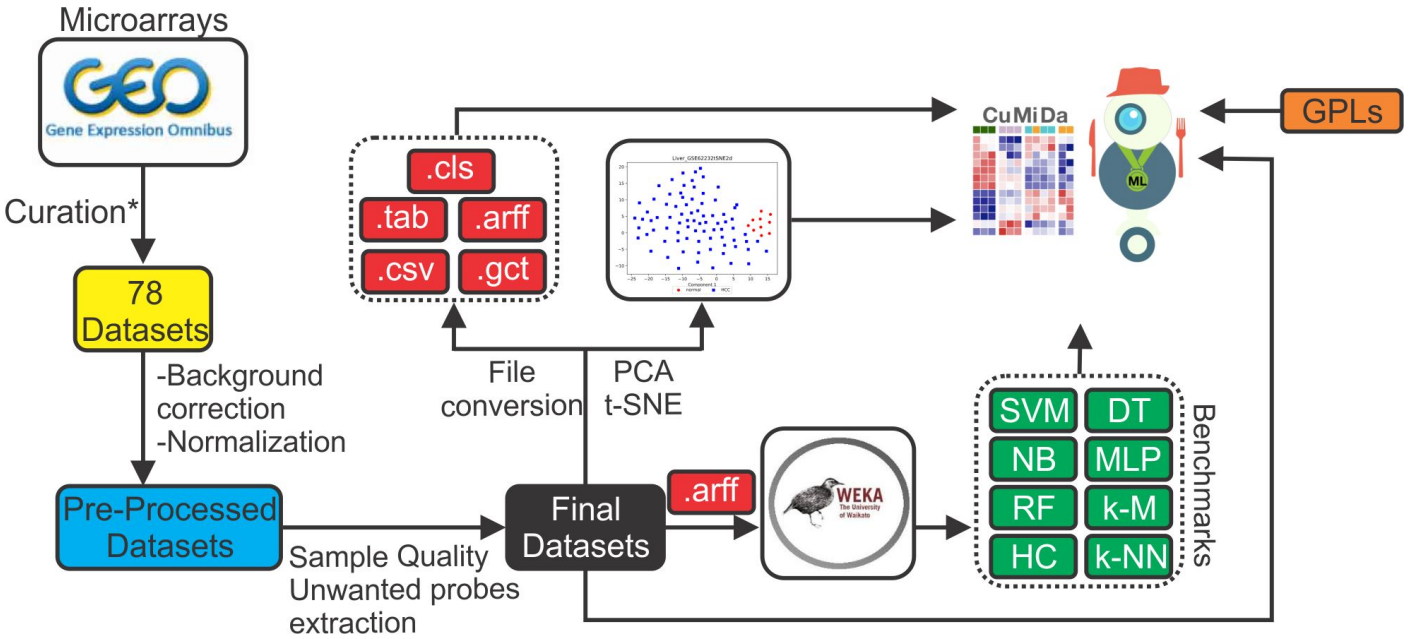


REAL MICROARRAY IMAGE

Mas na realidade...

	A	B	C	D	E	F	G	H	I	J	K	L
1	samples	type	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_at	1405_i_at	1431_at
2	GSM362958.CEL.gz	HCC	6.80119760710874	4.55318894640395	6.78779002909496	5.43089267353575	3.25022198170724	6.27268846398046	3.4134047464412	3.37490984706688	3.65411613599682	3.80498325183284
3	GSM362959.CEL.gz	HCC	7.58595616226449	4.19354041301431	3.76318259460317	6.00359349944804	3.3093867681399	6.29192656845271	3.7547770136409	3.58760263207689	5.13715938611725	8.6224754628958
4	GSM362960.CEL.gz	HCC	7.8033704814389	4.134075383015	3.43311303956077	5.39505709653068	3.47694393405601	5.82571255203611	3.50503558852534	3.68733289144417	4.51517535182232	12.6814391099332
5	GSM362964.CEL.gz	HCC	6.92084037613297	4.0006514934097	3.75450034120241	5.64529670292514	3.387530474805 89	6.47045799349257	3.62924920818033	3.57753407054759	5.19262423621383	11.7594120751621
6	GSM362965.CEL.gz	HCC	6.55648031523744	4.59901049029505	4.06615511575957	6.34453682825612	3.37208126056019	5.43928036335272	3.76221336675644	3.44071407349146	4.96162490605576	10.318552481299
7	GSM362966.CEL.gz	HCC	7.30668391006144	4.36869795997217	3.72489358089459	5.61296709092874	3.22458864145194	6.11220880672554	3.48660154139636	3.3149501822187	4.07110192135605	11.944573918641
8	GSM362970.CEL.gz	HCC	6.51292835592314	4.64456440565343	3.87271884659331	5.6865917726611	3.30420310841847	6.44245508824605	3.37885684430557	3.26577268196747	4.64010462565857	10.1967888008446
9	GSM362971.CEL.gz	HCC	6.82698754213955	4.01936601587141	4.12079159862386	5.71730950146647	3.21870382202531	6.69032703329845	3.77286897801043	3.51346182154437	5.88813125325068	12.5153547104115
10	GSM362972.CEL.gz	HCC	5.82078827848331	4.1594009925679	3.72900435654028	6.2171154352368	3.18383144008311	6.13059192981839	3.67742123833104	3.4016797078 56	4.78391990602284	12.6981102815037
11	GSM362976.CEL.gz	HCC	5.9275328265135	3.90591397539624	4.59021756466884	5.93549090963435	3.3647777499984	7.05783615012919	3.45456295354767	3.55231493561547	4.57788472236236	6.60509498565527
12	GSM362977.CEL.gz	HCC	5.62856178513269	3.92881995590034	4.8571091301983	6.30642665784121	3.27961837411023	6.54674268974066	3.68233421553639	3.57668134999204	4.22799221949975	3.748344598479 46
13	GSM362978.CEL.gz	HCC	6.09841757361098	4.45752780330972	3.99439026148466	5.73292905095257	3.36280682793552	6.97809176114277	3.47418272881083	3.50664240850734	4.20849010821785	4.29009126149271
14	GSM362982.CEL.gz	HCC	6.38762657207729	4.68799538667814	3.85478251355535	6.21822737708569	3.08222302712912	5.23572609261187	3.55217304909753	3.48640332061621	3.76379378225426	7.144559677397 36
15	GSM362984.CEL.gz	HCC	6.37522538336142	4.04805125867143	3.67044931930925	5.98933754265255	3.20246625382664	6.44163286878761	3.85477731752766	3.47225149425935	7.17069143839285	11.1891345283344
16	GSM362986.CEL.gz	HCC	6.93266499946296	4.60292142911026	4.21633575129576	5.58179928388985	3.02815457778632	6.20912434026108	3.35299034154755	3.34277749465531	5.46352397101956	12.7311657113786
17	GSM362988.CEL.gz	HCC	7.25535844960416	4.63563029909791	4.32854075845899	5.92269417415942	3.22199078758323	6.22190449941213	3.53584663738093	3.36400454681158	3.72975983671897	11.7634073358966
18	GSM362992.CEL.gz	HCC	6.16990686619275	4.55959083029603	3.44251412571729	6.05767087238335	3.22929083619527	5.9663905720069	3.63740970642988	3.35802913300437	4.32285374565323	8.296018454496 46
19	GSM362993.CEL.gz	HCC	8.702611285967	4.5727589417081	3.60329349161951	5.61303265328149	3.15232780042145	6.27772762939536	3.56383406210307	3.53064743723948	3.84231152408843	8.1659041300882
20	GSM362994.CEL.gz	HCC	6.76482163999378	4.28165273030982	3.88472119278071	6.20070231310914	3.1903797449068	6.49891778701107	3.46185227253752	3.31771948115732	5.54825462467011	12.5643275309247
21	GSM363008.CEL.gz	HCC	6.89173326997618	3.89934830146213	4.15418389037624	5.11307984280881	3.15619518152059	6.41914299664694	3.59398956916429	3.59780910067754	6.67426933341705	11.3200004786843
22	GSM363009.CEL.gz	HCC	9.24515459151105	5.52516401623998	4.2492737893893	5.32144021626249	3.14953271124594	5.75167306176937	3.5102232584056	3.49421175281972	4.62115010962122	10.640134987458
23	GSM363010.CEL.gz	HCC	10.2036809385542	4.38750392426001	4.86400382786457	6.04167547322937	3.22782381595483	4.89405457755297	3.54367134935043	3.9351253121344	4.61428213999159	11.6631860481393
24	GSM363011.CEL.gz	HCC	6.78216616252634	4.93379400808187	3.65656705689873	5.37289090421139	3.07899297401444	5.08403794010921	3.4711624344264	3.78480830512345	3.81215366577371	8.52870253684081
25	GSM363012.CEL.gz	HCC	6.01593272818864	4.94435214265907	4.2723560486957	5.76554771403007	3.1588937327206	5.88955933883493	3.50107439340929	3.62732206180902	4.56650861552184	12.609205710782
26	GSM363013.CEL.gz	HCC	7.4801122553537	4.48829575268765	4.20946718583123	5.16121006434855	3.20182366910594	6.2281552895264	3.67104654835145	3.25287752999281	5.53077999103044	7.38172245543906
27	GSM363015.CEL.gz	HCC	7.73372153457663	5.61893987311495	3.82750152732958	5.05102838028321	3.02710918909907	5.13790534683187	3.47191770098197	3.62029185943873	3.96904702528623	13.3174231335245
28	GSM363016.CEL.gz	HCC	5.86809915266974	4.24639962256086	3.86159690951172	5.83341771922282	3.14179061662545	6.87468444455053	3.68689513298409	3.427561976734	3.92303394246296	9.55136887242325
29	GSM363017.CEL.gz	HCC	6.28898429922968	5.30914607043119	4.10753938568554	5.40458017445522	3.08018468398627	6.5002554226958	3.9260921951267	3.4649327522747	4.106789173344	10.5394489705933
30	GSM363029.CEL.gz	HCC	6.84321505900151	4.65452808866373	4.58633394138622	4.94066602241705	3.31064895777859	6.4098394059198	3.2130662482615	3.57834438008277	5.73602278564337	11.4588677390233
31	GSM363030.CEL.gz	HCC	8.68638997709349	4.91966066842904	4.10221070510086	4.98228087608507	3.05775121546522	5.68917844090978	3.40367769809496	3.30041994003122	3.62365168080231	9.97755055922057
32	GSM363031.CEL.gz	HCC	5.73664136840904	4.32921107000053	3.99689832574993	4.92656687526748	3.14512776472711	6.36206320797895	3.90433661553501	3.53895072957314	4.70835494108223	10.0652927128643
33	GSM363032.CEL.gz	HCC	6.68190339487203	3.65907318228896	3.8776286480877	5.548796266581 1	3.12963120334136	6.43371571926667	3.77200741949821	3.67379686009843	6.20879328549831	11.8902326658131
34	GSM363033.CEL.gz	HCC	6.60768163260233	4.171716477834	3.84810776580581	5.01628590431952	3.15707645228569	6.01398600224565	3.67494302626603	3.49381996070804	5.76796336989196	5.62170775209241
35	GSM363034.CEL.gz	HCC	6.96794614857108	4.70058932127659	3.93045922921184	5.33699062327665	3.25378701235011	5.08553463505671	3.73650118154818	3.64212624955036	3.73654202426175	12.2657494635353
36	GSM363035.CEL.gz	HCC	6.62775948784058	3.86720135022778	3.92974189820237	5.81173457424161	3.32797019688142	6.42004136620377	3.8178044397589	3.54364806214802	4.8616909856137	11.9277627380745
37	GSM363036.CEL.gz	HCC	6.03010104997145	4.84860558050256	4.07228344735697	5.30249981394078	3.2793633737942	5.71557895384246	3.81729876757605	3.30793699305709	3.13495597655185	12.3562987481996
38	GSM363037.CEL.gz	HCC	5.96883440092929	5.0727098282114	4.01647778221287	5.11524930917977	3.20668790793104	5.83688576040371	3.50907190473833	3.59986724370337	3.2392695332282	12.1894741115421
39	GSM363038.CEL.gz	HCC	6.26981183983515	5.6542788111044	3.39836852448947	5.4906021963598	3.10952888595829	5.57165288595645	3.5194453586038	3.62498026973503	3.13393070619181	4.6227119271966
40	GSM363039.CEL.gz	HCC	6.61577369832654	4.18071000952337	3.78708038138071	5.73965445035121	3.1765494650974	6.01388465806638	3.71011840124224	3.74148308608491	3.54462167386714	13.342468129499
41	GSM363048.CEL.gz	HCC	6.32087223275949	4.62134779581375	3.83442769476652	5.15226544591636	2.95918093691951	5.7070378187334	3.29603266555585	3.56748716292217	3.70638276081107	8.88235247011474
42	GSM363049.CEL.gz	HCC	6.25292252358852	4.7186572153331	3.65393369535125	5.36538294244556	3.05976531190119	5.18867863033436	3.66146281150378	3.55644772906972	3.77369147527517	12.8649150415012
43	GSM363053.CEL.gz	HCC	6.59501074902191	5.37545270119183	3.67984966984489	5.14456193106467	3.18768453730654	5.85145392806153	3.47076425170495	4.1232396034402	3.83210369886728	9.04127206061433
44	GSM363055.CEL.gz	HCC	9.67749449464444	3.79206357783923	4.23465065185263	5.15307591816155	3.28243288981146	6.49057091608983	3.7454250442568	3.53619765154628	5.29949961461624	4.4065580305895
45	GSM363056.CEL.gz	HCC	8.14583616094427	5.58254840036409	4.24452775909892	5.39676018627489	3.06267439900211	5.85568054570108	3.67723811366602	3.54475606512553	3.3842863831052	
46	GSM363057.CEL.gz	HCC	6.17618317374308	3.87083017023233	7.23358608907206	5.17933205282205	3.08252431101478	5.71542025140122	3.40253690512998	3.67548330091057	3.3924772095767	11.8109131820272

CuMiDa


Microarrays
Gene Expression Omnibus
Curation*
78 Datasets
-Background correction
-Normalization
Pre-Processed Datasets
Sample Quality
Unwanted probes
extraction
.cls
.tab
.arff
.csv
.gct
File conversion
PCA
t-SNE
Final Datasets
.arff
WEKA
The University of Waikato
CuMiDa
ML
GPLs
SVM
DT
NB
MLP
RF
k-M
HC
k-NN
Benchmarks

Ang et al. (2015)

5.1 Dataset Analysis

Based on Tables 3, 4, and 5, the five most commonly used gene microarray expression datasets in the literatures are leukemia [67], colon [64], prostate [69], Diffuse Large B-Cell Lymphoma (DLBCL) [74], and Small round blue cell tumor (SRBCT) of childhood datasets [76]. The comparison of validation result using highest prediction accuracy based on CV and the number of selected genes in each literature are shown in Table 7, 8, 9, 10, and 11.

IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 13, NO. 5, SEPTEMBER/OCTOBER 2016 971

Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection

Jun Chin Ang, Andri Mirzal, Habibollah Haron, and Haza Nuzly Abdull Hamed

Abstract—Recently, feature selection and dimensionality reduction have become fundamental tools for many data mining tasks, especially for processing high-dimensional data such as gene expression microarray data. Gene expression microarray data comprises up to hundreds of thousands of features with relatively small sample size. Because learning algorithms usually do not work well with this kind of data, a challenge to reduce the data dimensionality arises. A huge number of gene selection are applied to select a subset of relevant features for model construction and to seek for better cancer classification performance. This paper presents the basic taxonomy of feature selection, and also reviews the state-of-the-art gene selection methods by grouping the literatures into three categories: supervised, unsupervised, and semi-supervised. The comparison of experimental results on top 5 representative gene expression datasets indicates that the classification accuracy of unsupervised and semi-supervised feature selection is competitive with supervised feature selection.

Index Terms—Feature selection, gene expression, semi-supervised, supervised, unsupervised

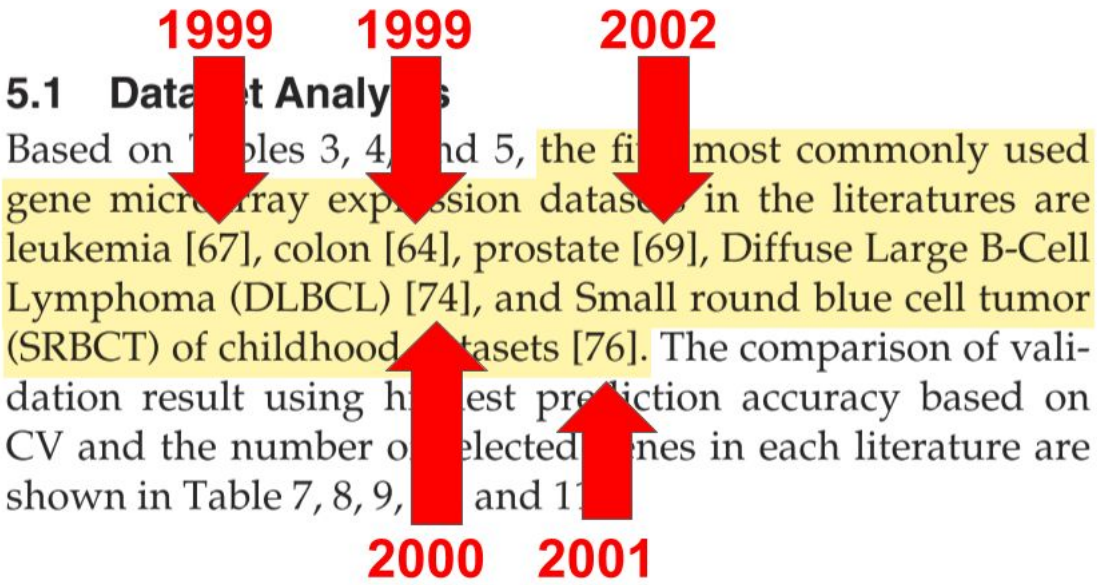
1 Introduction

Feature selection is a dimensionality reduction technique that is commonly used in the fields of machine learning, pattern recognition, statistics, and data mining. This technique aims to select a subset of relevant features from the original set of features according to some criteria. Some examples of feature selection techniques include Information Gain, Relief, Chi Squares, Fisher Score, and Lasso. Feature selection usually is used in the domains where the datasets comprise of thousands of features but with relatively small sample size (e.g., gene expression data). Feature selection that is applied to gene expression data is also known as gene selection [1]. Gene selection is necessary as the data usually contains many irrelevant, redundant, and noisy expressions, and also is effective for early tumor detection and cancer discovery as it leads to a more reliable cancer diagnosis or prognosis and better clinical treatment [2].

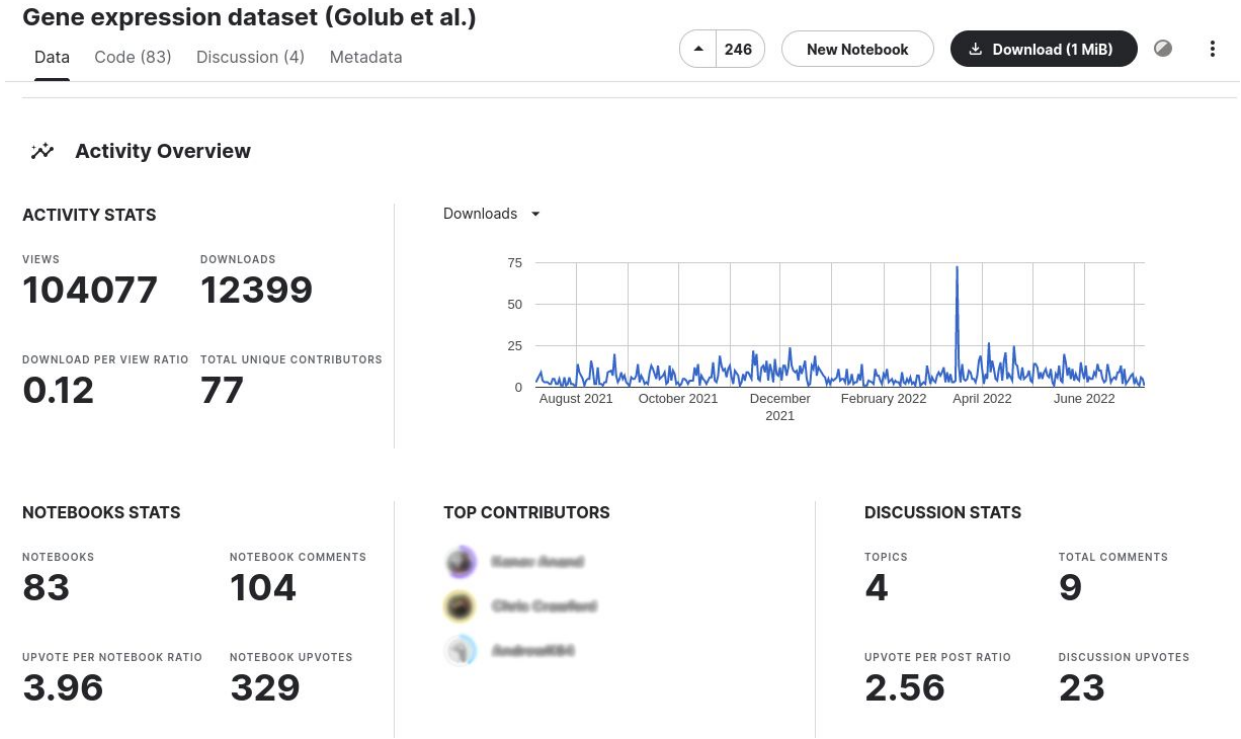
The gene expression data can either be fully labeled, unlabeled, or partially labeled. This leads to the development of supervised, unsupervised and semi-supervised gene selection to discover the biological patterns and class prediction [3]. Typically, unlabeled data is composed of samples and their features without the presence of labels process of selecting a feature subset based on some criteria for measuring importance and relevance of the features by using labeled data. Unsupervised feature selection, where there is no prior knowledge about the true functional classes, evaluates feature relevance by exploiting the innate structures of the data, such as data variance, separability, and data distribution. A semi-supervised feature selection integrates a small amount of labeled data into unlabeled data as additional information to improve the performance of an unsupervised feature selection. Even though there are a lot of review papers on gene selection in the literatures [4], [5], [6], [7], [8], but to the best of our knowledge, there is no detailed discussion on the methods and separates them in such three categories as in this paper.

This paper is divided into six sections. Section 2 presents an overview of feature selection and Section 3 gives a review on some gene selection approaches especially that have been proposed over the past five years and further group them in three categories: supervised, unsupervised and semi-supervised approaches. Section 4 describes the challenges inherent in gene selection task. Section 5 discusses the gene selection approaches reviewed in the previous sections and collates the experimental results of gene

10.1109/TCBB.2015.2478454


1999
1999
2002
5.1 Data t Analy s
Based on bles 3, 4, nd 5, the fi most commonly used
gene micr ray exp ssion datas in the literatures are
leukemia [67], colon [64], prostate [69], Diffuse Large B-Cell
Lymphoma (DLBCL) [74], and Small round blue cell tumor
(SRBCT) of childhood tasets [76]. The comparison of vali-
dation result using h est pr iction accuracy based on
CV and the number o elected nes in each literature are
shown in Table 7, 8, 9, and 1
2000
2001

Kaggle



https://www.kaggle.com/datasets/crawford/gene-expression/data

Bolón-Canedo et al. (2014)

Information Sciences 282 (2014) 111–135

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A review of microarray datasets and applied feature selection methods

V. Bolón-Canedo [a,*], N. Sánchez-Maroño [a], A. Alonso-Betanzos [a], J.M. Benítez [b], F. Herrera [b,c]

Table 5
Dataset description for binary datasets: *s* and *f* are the number of samples and features, respectively.

Dataset	*s*	*f*	Distribution	Original Ref.	Year	Where used	Download
B_MD	34	7129	26–74%	[95]	2002	[101]	*unknown*
Bone Lesion	173	12,625	*unknown*	[117]	2003	[101]	*unknown*
Brain	21	12,625	33–67%	[88]	2003	[27,120]	[2]
Brain_Tumor1	60	7129	*unknown*	*unknown*	*unknown*	[81]	*unknown*
Brain_Tumor2	50	12,625	*unknown*	*unknown*	*unknown*	[81]	*unknown*
Breast	22	3226	*unknown*	[58]	2001	[71]	*unknown*
Breast Cancer	97	24,481	*unknown*	[118]	2002	[24,23,27]	[8]
Breast-test	19	24,481	37–63%	[118]	2002	[23,48]	[8]
Breast-train	78	24,481	56–44%	[118]	2002	[23,48]	[8]
BreastER	49	7129	49–51%	[123]	2001	[101]	*unknown*
BR-ER49	49	6817	49–51%	[123]	2001	[73]	*unknown*
C_MD	60	7129	35–65%	[95]	2002	[101]	*unknown*
Celiac	132	22,185	*unknown*	[57]	2009	[101]	*unknown*
CNS/Embrional-T	60	7129	35–75%	[95]	2002	[124,24,25,23,27,105,125]	[2,3]
Colon	62	2000	35–65%	[14]	1999	[39,101,78,107,80,71,124,11,24,23,27,73,85,48,122,108,31,125,120]	[2,7,8,3]
Colon-epi	202	44,290	*unknown*	[87]	2008	[101]	*unknown*
DLBCL	77	5470	75–25%	[104]	2002	[39,33,122,125]	[110]
DLBCL	47	4026	49–51%	[13]	2000	[71,24,25,23,27,105]	[8,2]
DLBCL	77	7129	75–25%	[104]	2002	[124,81]	[2,9]
GLI-85	85	22,283	31–69%	[40]	2004	[27]	[5]
Leukemia/ ALLAML	72	7129	35–65%	[46]	1999	[39,101,107,71,124,11,86,24,25,27,73,85,31,120]	[2,8]
Leukemia_test	34	7129	71–29%	[46]	1999	[119,23,48,108]	[3,2,8]
Leukemia_train	38	7129	59–41%	[46]	1999	[119,23,48,108]	[3,2,8]
Lung	52	918	75–25%	[43]	2001	[101]	*unknown*
Lung	181	12,533	83–17%	[49]	2002	[24,25,27,48]	[8,2]
Lung	410	2428	34–66%	[103]	2008	[31]	*unknown*
Lung_test	149	12,533	90–10%	[49]	2002	[23,48]	[8,2]
Lung_train	32	12,533	50–50%	[49]	2002	[23,48]	[8,2]
LUNG181	181	12,600	17–83%	[49]	2002	[73]	*unknown*
LYM77	77	6817	25–75%	[104]	2002	[73]	[8]
Lymphoma/B-cell1	45	4026	49–51%	[13]	2000	[11,108]	[2]
Moffitt colon cancer	122	2619	31–69%	[38]	2005	[31]	*unknown*
Ovarian	253	15,154	36–64%	[93]	2002	[24,25,23,27,120,107]	[8]
Prostate	102	6033	51–49%	[106]	2002	[78,11]	*unknown*
Prostate	136	12,600	43–57%	[106]	2002	[124,24,25,27]	[2,8]
Prostate-test	34	12,600	26–74%	[106]	2002	[102,85,23,48]	[8]
Prostate-train	102	12,600	49–51%	[106]	2002	[102,86,73,85,81,23,48,105]	[8,9]
Prostate Tumor	102	10,509	51–49%	[106]	2002	[39,33,125,120]	[110]
SMK-CAN-187	187	19,993	48–52%	[109]	2007	[27,119,108]	[5]

Table 6
Dataset description for multiclass datasets: *s*, *f* and *c* are the number of samples, features and classes, respectively.

Dataset	*s*	*f*	*c*	Original Ref.	Year	Where used	Download
9-Tumors	60	5726	9	[111]	2001	[39,33,81,125]	[110,9]
11-Tumors/Carcinomas	174	12,533	11	[114]	2001	[39,33,86,81,125]	[110,9]
14-Tumors	308	15,009	26	[96]	2001	[39,33,69,81,125]	[110,9]
Brain Tumor 1	90	5920	5	[95]	2002	[39,33,81,125]	[110,9]
Brain Tumor 2	50	10,367	4	[88]	2003	[39,33,81]	[110,9]
CLL-SUB-111	111	11,340	3	[55]	2004	[119,108]	[5]
GCM	198	16,306	14	[96]	2001	[71,24,23,27]	[2]
GCM	190	16,063	14	[96]	2001	[125]	[3]
GLA-BRA-180	180	49,151	4	[115]	2006	[27,108]	[5]
Glioma	50	12,625	4	[88]	2003	[86]	[2]
Global Cancer Map/GCM-Train	144	16,063	14	[96]	2001	[119,108]	[3]
Leukaemia	110	22,278	*unknown*	*unknown*	*unknown*	[16]	*unknown*
Leukemia 1	72	5327	3	[46]	1999	[39,33,81,125]	[110,9]
Leukemia 2	72	11,225	3	[17]	2002	[39,33,81,125]	[110,9]
Lung	254	8359	5	[20]	2001	[113]	[2]
Lung-Cancer	203	12,601	5	[20]	2001	[39,33,86,81,125]	[110,9]
Lymphoma/B-cell3	96	4026	9	[13]	2000	[39,80,24,23,27,108,120]	[110]
MLL	72	8359	3	[17]	2002	[113]	[2]
MLL-train	57	12,582	3	[17]	2002	[102]	[8]
MLL-test	15	12,582	3	[17]	2002	[102]	[8]
TOX-171	171	5748	4	[112]	2010	[27,119,108]	[5]
SRBCT	83	2309	4	[66]	2001	[39,113,33,71,81]	[110,9]
SRBCT-train	63	2309	4	[66]	2001	[102]	[110]
SRBCT-test	20	2309	4	[66]	2001	[102]	[110]
Yeast	79	2467	*unknown*	[30]	2000	[107]	*unknown*

https://doi.org/10.1016/j.ins.2014.05.042

Bolón-Canedo et al. (2014)

4.1. DNA microarray repositories

Although in the initial development of DNA microarray data analysis it was difficult to find datasets to deal with, in recent years there has been a growing number of public microarray data repositories of a wide spectrum of cancer types available for the scientific community. The most famous are listed below:

- *ArrayExpress*, from the European Bioinformatics Institute [1]:
  http://www.ebi.ac.uk/arrayexpress/

Table 4
Other feature selection methods used on microarray data. Type of evaluation (ranker/subset) and type of data (binary/multiclass).

Method	Original Ref.	Type (r/s)	Data (b/m)
CFS-TGA	[33]	s	m
E1-cp	[25]	s	b
E1-ni	[25]	s	b
E1-ns	[25]	s	b
E2	[25]	s	b
Ensemble RFE	[11]	s	b
EF	[24]	s	m
FAST	[108]	s	m
GADP	[71]	s	m
GC	[78]	r	b
MCF-RFE	[124]	s	b
MFMW	[73]	s	b
R-m-GA	[105]	s	b
SRF	[125]	s	m
SVM-RFE with MRMR	[85]	r	b

- *Cancer Program Data Sets*, from the Broad Institute [2]:
  http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi
- *Dataset Repository*, from the Bioinformatics Research Group of Universidad Pablo de Olavide [3]:
  http://www.upo.es/eps/bigs/datasets.html
- *Feature Selection Datasets*, from Arizona State University [5]:
  http://featureselection.asu.edu/datasets.php
- *Gene Expression Model Selector*, from Vanderbilt University [110]:
  http://www.gems-system.org
- *Gene Expression Omnibus*, from the National Institutes of Health [6]:
  http://www.ncbi.nlm.nih.gov/geo/
- *Gene Expression Project*, from Princeton University [7]:
  http://genomics-pubs.princeton.edu/oncology/
- *Kent Ridge Bio-Medical Dataset Repository*, from the Agency for Sciency, Technology and Research [8]:
  http://datam.i2r.a-star.edu.sg/datasets/krbd
- *Stanford Microarray Database*, from Stanford University [10]:
  http://smd.stanford.edu/

Problemas com dados

- **Northcutt et al. (2021):** 3,4% dos rótulos nos 10 *datasets* mais usados em visão computacional, linguagem natural e áudio contêm erros.

- **Roberts et al. (2021):** nenhum dos modelos treinados para diagnóstico de COVID-19 a partir de radiografia ou tomografia do tórax pode ser usado na prática clínica por causa de falhas metodológicas, incluindo vieses nos conjuntos de dados.

- **Koch et al. (2021):** um dos problemas atuais em aprendizado de máquina são estudos de *benchmarking* usando dados para tarefas para as quais eles não foram originalmente desenhados.

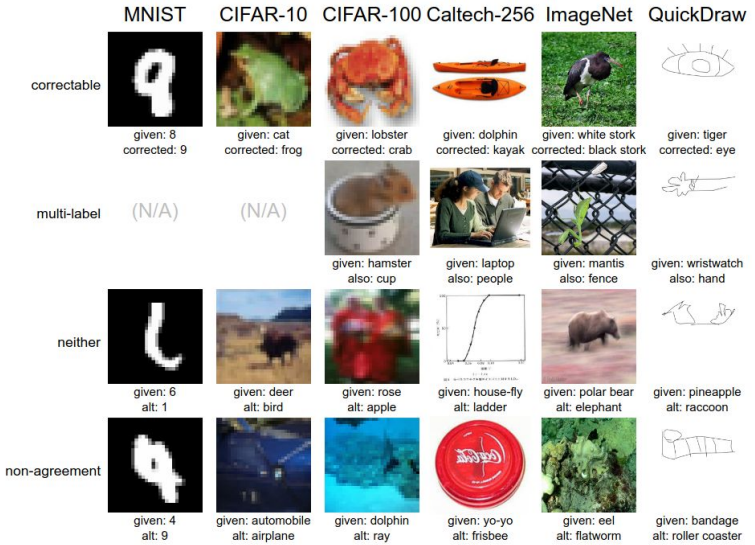
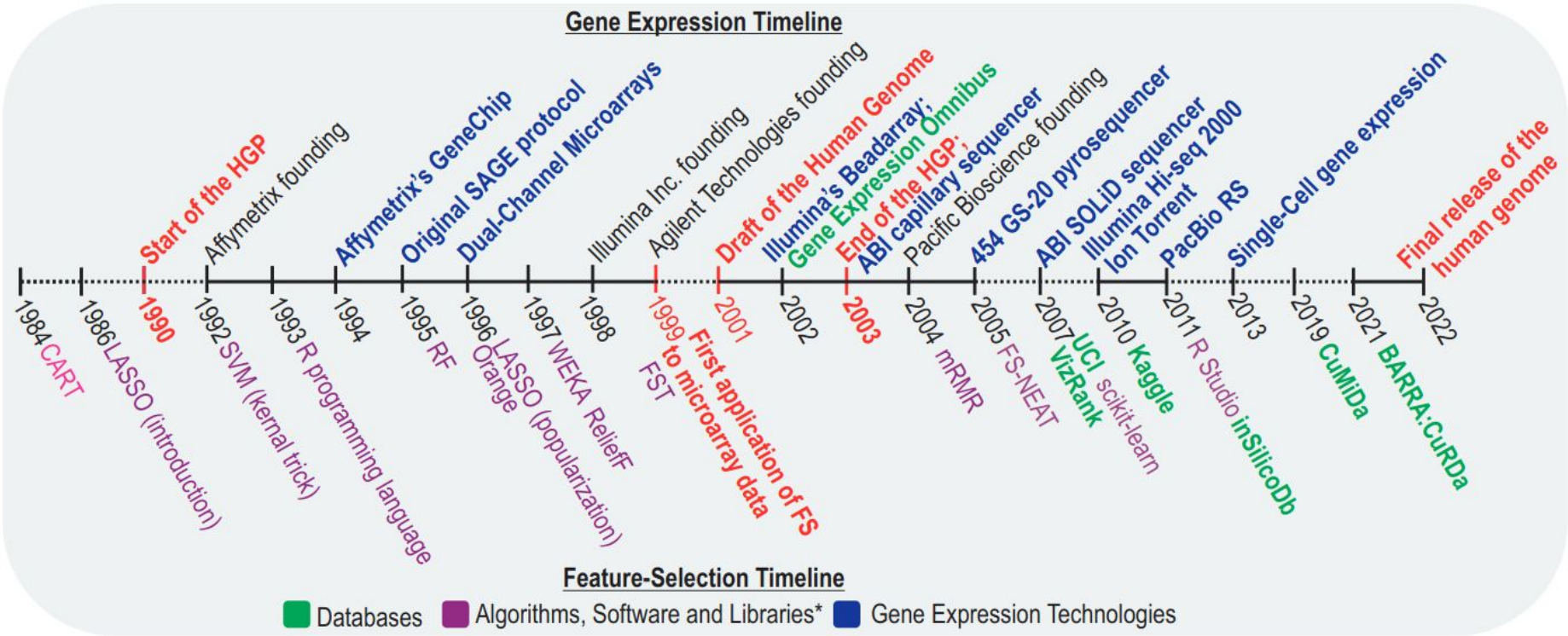


Figure 1: An example label error from each category (Section 4) for image datasets. The figure shows given labels, human-validated corrected labels, also the second label for multi-class data points, and CL-guessed alternatives. A gallery of label errors across all 10 datasets, including text and audio datasets, is available at https://labelerrors.com.

arXiv:2103.14749


Gene Expression Timeline
1984 CART
1986 LASSO (introduction)
1990 Start of the HGP
1992 SVM (kernal trick)
1993 R programming language
1994 Affymetrix founding
1995 RF
1996 LASSO (popularization) Orange
1997 WEKA ReliefF
1998 Affymetrix's GeneChip
Original SAGE protocol
Dual-Channel Microarrays
1999 First application of FS to microarray data FST
2001 Illumina Inc. founding
Agilent Technologies founding
2002 Draft of the Human Genome
Illumina's Beadarray;
Gene Expression Omnibus
2003 End of the HGP;
ABI capillary sequencer
2004 mRMR
2005 FS-NEAT
Pacific Bioscience founding
454 GS-20 pyrosequencer
2007 UCI scikit-learn
ABI SOLiD sequencer
VizRank
2010 Kaggle
Illumina Hi-seq 2000
Ion Torrent
2011 R Studio inSilicoDb
PacBio RS
2013 Single-Cell gene expression
2019 CuMiDa
2021 BARRA:CuRDa
2022 Final release of the human genome
Feature-Selection Timeline
Databases
Algorithms, Software and Libraries*
Gene Expression Technologies

Vamos à revisão

Bases de artigos

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Supplementary Material 2: The use of gene expression datasets in feature selection research: 20 years of inherent bias?**												
2	**Bruno Iochins Grisci, Bruno César Feltes, Joice de Faria Poloni, Pedro Henrique Narloch, and Márcio Dorn**												
3	**S1-Table PUBLICATIONS**												
4	**ID**	**Source**	**Search1**	**Search2**	**Title**	**Authors**	**DOI**	**Year**	**Publisher**	**Volume**	**Issue**	**Pages**	**PMID**
5	P0007	WOS	RNA-SEQ	FEATURE SELECTIO	Omic and Electronic Health Record Big Data Analytics for Precision Medicine	Wu, PY; Che	10.1109/TB	2017	IEEE TRANS	64	2	263-273	277404
6	P0010	PUBMED	MICROARRA	FEATURE SELECTIO	A BAYESIAN NONPARAMETRIC MIXTURE MODEL FOR SELECTING GENES AND GENE SUB	Zhao Y, Kang	10.1214/14	2014	Ann Appl Sta	8	2	999-1021	259842
7	P0011	IEEE	MICROARRA	FEATURE SELECTIO	A binary PSO feature selection algorithm for gene expression data	S. Dara; H. B	10.1109/EIC	2014	2014 Intern	NA	NA	1-6	NA
8	P0012	IEEE	MICROARRA	FEATURE SELECTIO	A Biologically Verified Classification of Microarray Data	R. Mondal; E	10.1109/CIC	2014	2014 Intern	NA	NA	686-690	NA
9	P0013	SCOPUS	MICROARRA	FEATURE SELECTIO	A biology-driven approach identifies the hypoxia gene signature as a predictor of the outcc	Fardin P., Ba	10.1186/14	2010	Molecular C	9	NA	NA	206242
10	P0015	PUBMED	MICROARRA	FEATURE SELECTIO	A CBR framework with gradient boosting based feature selection for lung cancer subtype c	Ramos-Gonz	10.1016/j.cc	2017	Comput Biol	86	NA	98-106	285273
11	P0016	PUBMED	MICROARRA	FEATURE SELECTIO	A centroid-based gene selection method for microarray data classification.	Guo S, Guo D	10.1016/j.jt	2016	J Theor Biol	400	NA	32-41	270567
12	P0017	PUBMED	MICROARRA	FEATURE SELECTIO	A class imbalance-aware Relief algorithm for the classification of tumors using microarray	He Y, Zhou J,	10.1016/j.cc	2019	Comput Biol	80	NA	121-127	309470
13	P0018	IEEE	RNA-SEQ	FEATURE SELECTIO	A Class-Information-Based Sparse Component Analysis Method to Identify Differentially E	J. Liu; Y. Xu;	10.1109/TC	2016	IEEE/ACM T	13	2	392-398	NA
14	P0020	SCOPUS	MICROARRA	FEATURE SELECTIO	A classification model on tumor cancer disease based mutual information and firefly algori	Jabbar S.F.	10.21533/p	2019	Periodicals (	7	3	1152-1162	NA
15	P0021	SCOPUS	MICROARRA	FEATURE SELECTIO	A clustering approach for feature selection in microarray data classification using random	Aydadenta H	10.3745/JIP	2018	Journal of In	14	5	1167-1175	NA
16	P0023	SCOPUS	MICROARRA	FEATURE SELECTIO	A clustering-based method for gene selection to classify tissue samples in lung cancer	Castellanos-	10.1007/97	2016	Advances in	477	NA	99-107	NA
17	P0024	SCOPUS	MICROARRA	FEATURE SELECTIO	A combining dimensionality reduction approach for cancer classification	Han L., Zhou	10.1007/97	2015	Lecture Note	9426	NA	340-347	NA
18	P0025	PUBMED	MICROARRA	FEATURE SELECTIO	A comparative analysis of biomarker selection techniques.	Dessì N, Pas	10.1155/20	2013	Biomed Res	2013	NA	387673	243249
19	P0026	IEEE	MICROARRA	FEATURE SELECTIO	A comparative analysis of feature selection algorithms on classification of gene microarray	J. Jeyachidra	10.1109/ICI	2013	2013 Intern	NA	NA	1088-1093	NA
20	P0027	SCOPUS	MICROARRA	FEATURE SELECTIO	A comparative analysis of feature selection methods for biomarker discovery in study of to	Zhang X., Jor	10.1007/97	2019	Communica	1056 CCIS	NA	114-123	NA
21	P0028	PUBMED	MICROARRA	FEATURE SELECTIO	A comparative analysis of swarm intelligence techniques for feature selection in cancer cla	Gunavathi C	10.1155/20	2014	ScientificWo	2014	NA	693831	251573
22	P0029	IEEE	MICROARRA	FEATURE SELECTIO	A Comparative Performance Evaluation of Random Forest Feature Selection on Classificati	M. A. Latief;	10.1109/ICI	2019	2019 3rd Int	NA	NA	1-6	NA
23	P0030	SCOPUS	MICROARRA	FEATURE SELECTIO	A Comparative Performance Evaluation of Supervised Feature Selection Algorithms on Mic	Arun Kumar	10.1016/j.pi	2017	Procedia Co	115	NA	209-217	NA
24	P0032	SCOPUS	MICROARRA	FEATURE SELECTIO	A Comparative study and analysis of data mining classifiers for microarray based cancer pa	Subasree S.,	10.1145/29	2016	ACM Interna	25-26-Augu	NA	NA	NA
25	P0033	SCOPUS	MICROARRA	FEATURE SELECTIO	A comparative study of feature selection and classification methods for gene expression da	Abusamra H	10.1016/j.pi	2013	Procedia Co	23	NA	mai./14	NA
26	P0034	SCOPUS	MICROARRA	FEATURE SELECTIO	A comparative study of feature selection and classification techniques for high-throughput	Alkuhlani A.	10.1007/97	2017	Advances in	533	NA	793-803	NA
27	P0035	IEEE	MICROARRA	FEATURE SELECTIO	A Comparative Study of Feature Selection Methods on Genomic Datasets	J. Rahimipou	10.1109/CB	2019	2019 IEEE 3	NA	NA	471-476	NA
28	P0036	IEEE	MICROARRA	FEATURE SELECTIO	A comparative study of gene selection methods for cancer classification using microarray c	M. Babu; K. S	10.1109/ICI	2016	2016 Secon	NA	NA	204-211	NA
29	P0037	SCOPUS	MICROARRA	FEATURE SELECTIO	A Comparative Study of Gene Selection Methods for Microarray Cancer Classification	Alshamlan H	10.1007/97	2019	Lecture Note	520	NA	585-595	NA
30	P0038	SCOPUS	MICROARRA	FEATURE SELECTIO	A Comparative Study of Improvements Filter Methods Bring on Feature Selection Using Mi	Wang Y., Fan	10.1007/97	2014	Lecture Note	8423 LNCS	NA	55-62	NA
31	P0039	PUBMED	MICROARRA	FEATURE SELECTIO	A comparative study of improvements Pre-filter methods bring on feature selection using r	Wang Y, Fan	10.1186/20	2014	Health Inf Sc	2	NA	7	258256
32	P0040	SCOPUS	MICROARRA	FEATURE SELECTIO	A comparative study of multiclass feature selection on RNAseq and microarray data	Zhang S., Wa	10.1504/IIC	2019	Internationa	12	2	128-142	NA

https://doi.org/10.1002/widm.1523

Supplementary Material 2: The use of gene expression datasets in feature selection research: 20 years of inherent bias?

Bruno Iochins Grisci, Bruno César Feltes, Joice de Faria Poloni, Pedro Henrique Narloch, and Márcio Dorn

S2-Table DATASETS

DATASETS						
ID DATA	DATASET	YEAR	METHOD	SPECIES	BACKGROUND	Link Dataset	AKA
665	Marti et al. (2015)	2015	*Microarray*	***Equus ferus***	Allergies	https://www.sciencedirect.com/science/article/pii/S0165242715001397?via%3Dihub#se	
202	Arisi et al. (2011)	2011	*Microarray*	***Mus musculus***	Alzheimer's Disease	https://content.iospress.com/articles/journal-of-alzheimers-disease/jad101881	
795	GSE11882	2008	*Microarray*	*Homo sapiens*	Alzheimer's Disease	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11882	Berchtold et
258	GSE1297	2004	*Microarray*	*Homo sapiens*	Alzheimer's Disease	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1297	Blalock et al
259	GSE5281	2006	*Microarray*	*Homo sapiens*	Alzheimer's Disease	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5281	
260	Ray et al. (2007)	2007	*Microarray*	*Homo sapiens*	Alzheimer's Disease	https://www.nature.com/articles/nm1653#MOESM5	
441	E-TABM-940	2010	*Microarray*	*Homo sapiens*	Amyotrophic lateral sclerosis	https://www.ebi.ac.uk/arrayexpress/experiments/E-TABM-940/	
100	Yasvoina et al. (2013)	2013	**RNA-Seq**	***Mus musculus***	Amyotrophic lateral sclerosis	https://www.jneurosci.org/content/33/18/7890	
546	GSE83091	2016	*Microarray*	*Homo sapiens*	Appendicitis	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE83091	Chawla et a
350	GSE20129	2010	*Microarray*	*Homo sapiens*	Atherosclerosis	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20129	
379	GSE23746	2010	*Microarray*	*Homo sapiens*	Atherosclerosis	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23746	
156	GSE25507	2010	*Microarray*	*Homo sapiens*	Autism	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25507	
230	E-MTAB-37	2008	*Microarray*	*Homo sapiens*	Biology of multiple cancers	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-37/	
349	GSE2564	2005	*Microarray*	*Homo sapiens*	Biology of multiple cancers	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2564	Getz et al. (
595	GSE32474	2011	*Microarray*	*Homo sapiens*	Biology of multiple cancers	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32474	Reinhold et
606	GSE52582	2013	*Microarray*	*Homo sapiens*	Biology of multiple cancers	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52582	Stafford et a
302	GSE5364	2006	*Microarray*	*Homo sapiens*	Biology of multiple cancers	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5364	
191	GSE68086	2015	**RNA-Seq**	*Homo sapiens*	Biology of multiple cancers	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68086	
166	GSE74251	2015	**RNA-Seq**	*Homo sapiens*	Biology of multiple cancers	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74251	
44	Pan-Cancer dataset	**-1**			Biology of multiple cancers		
791	Pfister et al. (2009)	2009	*Microarray*	*Homo sapiens*	Biology of multiple cancers	https://mct.aacrjournals.org/content/8/7/1878	
66	Ramaswamy et al. (2001)	2001	*Microarray*	*Homo sapiens*	Biology of multiple cancers	https://www.pnas.org/content/98/26/15149.short	
84	Ross et al. (2000)	2000	*Microarray*	*Homo sapiens*	Biology of multiple cancers	https://pubmed.ncbi.nlm.nih.gov/10700174/	
187	Staunton et al. (2001)	2001	*Microarray*	*Homo sapiens*	Biology of multiple cancers	https://www.pnas.org/content/pnas/98/19/10787.full.pdf	
30	TCGA	**-1**	**RNA-Seq**	*Homo sapiens*	Biology of multiple cancers	http://www.cbioportal.org/study/summary?id=brca_tcga_pub2015	

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Supplementary M**	**laterial 2: The use of gene expression datasets in feature selection research: 20 years of inherent bias?**											
2	**Bruno Iochins Gr**	**isci, Bruno César Feltes, Joice de Faria Poloni, Pedro Henrique Narloch, and Márcio Dorn**											
3	**S3-Table LINKS**												
4		Yellow if indirectly cited											
5	**PAPER ID**	**DATASET**	**DATASET**	**DATASET**	**DATASET**	**DATASET**	**DATASET**	**DATASET**	**DATASET**	**DATASET**	**DATASET**	**DATASET**	**DATASET**	**DATA**
6	P0020	Broken Link	Alon (1999)											
7	P0026	Broken Link	Alon (1999)											
8	P0032	Dyrskjot et al. (2	GSE349_350	GSE3726	Missing	Golub et al. (1	Bhattacharjee	Khan et al. (20	GSE2443					
9	P0038	GSE36961	GSE36964											
10	P0044	Golub et al. (19												
11	P0050	GSE15471												
12	P0056	Golub et al. (19	Alon (1999)	Broken Link	Pomeroy et al.									
13	P0062	Missing	Golub et al. (1	Alon (1999)	Khan et al. (20									
14	P0068	Golub et al. (19	Khan et al. (20	Armstrong et a										
15	P0080	GSE54460												
16	P0086	Golub et al. (19	Singh et al. (2C	Shipp et al. (2C	Veer et al. (20(	Bhattacharjee	Kong et al. (20	Alon (1999)	Wu et al. (200	Boyle et al. (2C	Zhang et al. (2(	Smith et al. (2C		
17	P0092	GSE54460												
18	P0098	Broken Link	Golub et al. (1	Khan et al. (20	Alon (1999)	Singh et al. (2C	Veer et al. (20(							
19	P0104	Broken Link	Golub et al. (1	Khan et al. (20	Shipp et al. (2C	Alon (1999)								
20	P0122	Pollen et al. (20												
21	P0128	Golub et al. (19	Alon (1999)	Khan et al. (20	Shipp et al. (2C	Armstrong et a	Pomeroy et al.							
22	P0134	Golub et al. (19	Alon (1999)	Khan et al. (20	Armstrong et a	Singh et al. (2C								
23	P0140	Simulation	Authors Datas											
24	P0146	Golub et al. (19	Pomeroy et al.											
25	P0152	Simulation	TCGA											
26	P0158	Golub et al. (19	Pomeroy et al.											
27	P0164	Simulation	Broken Link	Golub et al. (1	Armstrong et a	Pomeroy et al.	Khan et al. (20							
28	P0170	Alon (1999)	Khan et al. (20	Singh et al. (2C	Shipp et al. (2C	Armstrong et a	Bhattacharjee							
29	P0176	Shipp et al. (200	Golub et al. (1	Singh et al. (2C										
30	P0182	Missing	Shipp et al. (2C	Khan et al. (20	Singh et al. (2C	Bhattacharjee	Pomeroy et al.							

	All	IEEE Xplore	PubMed	WOS	Scopus
Number of publications	1284	375	315	47	547
Mean number of datasets per publication	3.94 ± 3.38	3.76 ± 3.17	3.66 ± 3.47	3.66 ± 3.03	4.26 ± 3.48
Median number of datasets per publication	3	3	2	3	3
Min number of datasets per publication	1	1	1	1	1
Max number of datasets per publication	27	27	25	12	27
Indirect citations[a]	32.3%	39.4%	15.5%	23.4%	37.8%
Broken links[b]	7.7%	8.5%	5.3%	10.6%	8.4%
Missing reference[c]	5.1%	9.6%	0.3%	6.3%	4.7%
Simulated data[d]	6.5%	8.0%	8.5%	6.3%	4.3%
Author's dataset[e]	4.6%	2.1%	9.2%	4.2%	3.8%

[a]Articles citing an intermediary study instead of the original source of the data.
[b]Articles containing URLs that do not work or are not up-to-date.
[c]Articles without proper references for the used datasets.
[d]Articles using data from computational simulations and not from biological experiments.
[e]Articles using their own datasets.

Citação indireta

4. Datasets

To demonstrate the efficiency of our method, the proposed method is evaluated on five gene expression datasets used in the literature.

(1) Colon dataset: colon dataset [15] is derived from colon cancer patient samples. It consists of the expression levels of 1909 genes of 62 patients among which 40 are colon cancer cases and 22 are normal cases.

(2) Ovarian dataset: ovarian dataset [16] often serves as benchmark for microarray data analysis in most literatures. The dataset provided here includes 91 controls (normal) and 162 ovarian cancers. There are a total of 15,154 genes.

(3) CNS dataset: CNS dataset [15] contains 60 patient samples out of which 39 are normal cases and 21 are cancer cases. There are 7129 genes in the dataset.

(4) Leukemia dataset: leukemia dataset [6] is another dataset widely used in the literature, which is taken as our benchmark dataset. The leukemia dataset contains the expression levels of 7129 genes taken from 72 samples. Labels indicate that there are 47 cancer cases and 25 normal cases.

(5) Breast dataset: breast dataset [17] consists of 97 samples, of which 51 are normal cases and 46 are cancer cases. Large number of features against small sample size is the trademark of the breast cancer dataset. Table 2 shows the summary of these datasets.

5. Guyon I., Weston J., Barnhill S., and Vapnik V., Gene selection for cancer classification using support vector machines, *Machine Learning*. (2002) **46**, no. 1–3, 389–422, https://doi.org/10.1023/A:1012487302797, 2-s2.0-0036161259.

View | Web of Science® | Google Scholar

6. Furey T., Cristianini N., Duffy N., Bednarski D. W., Schummer M., and Haussler D., Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*. (2000) **16**, no. 10, 906–914, https://doi.org/10.1093/bioinformatics/16.10.906, 2-s2.0-0033636139.

View | CAS | PubMed | Web of Science® | Google Scholar

7. Statnikov A., Aliferis C. F., Tsamardinos I., Hardin D., and Levy S., A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, *Bioinformatics*. (2005) **21**, no.

Links quebrados

www.c2i.ntu.edu.sg

A. Data Acquisition

For this study, the breast cancer microarray data is obtained from the Centre for Computational Intelligence (*http://www.c2i.ntu.edu.sg/*). The dataset, which contains 97 instances with 24461 genes and 2 classes, is in the Artificial-Relation File Format, ARFF.

Hmm. We're having trouble finding that site.

We can't connect to the server at www.c2i.ntu.edu.sg.

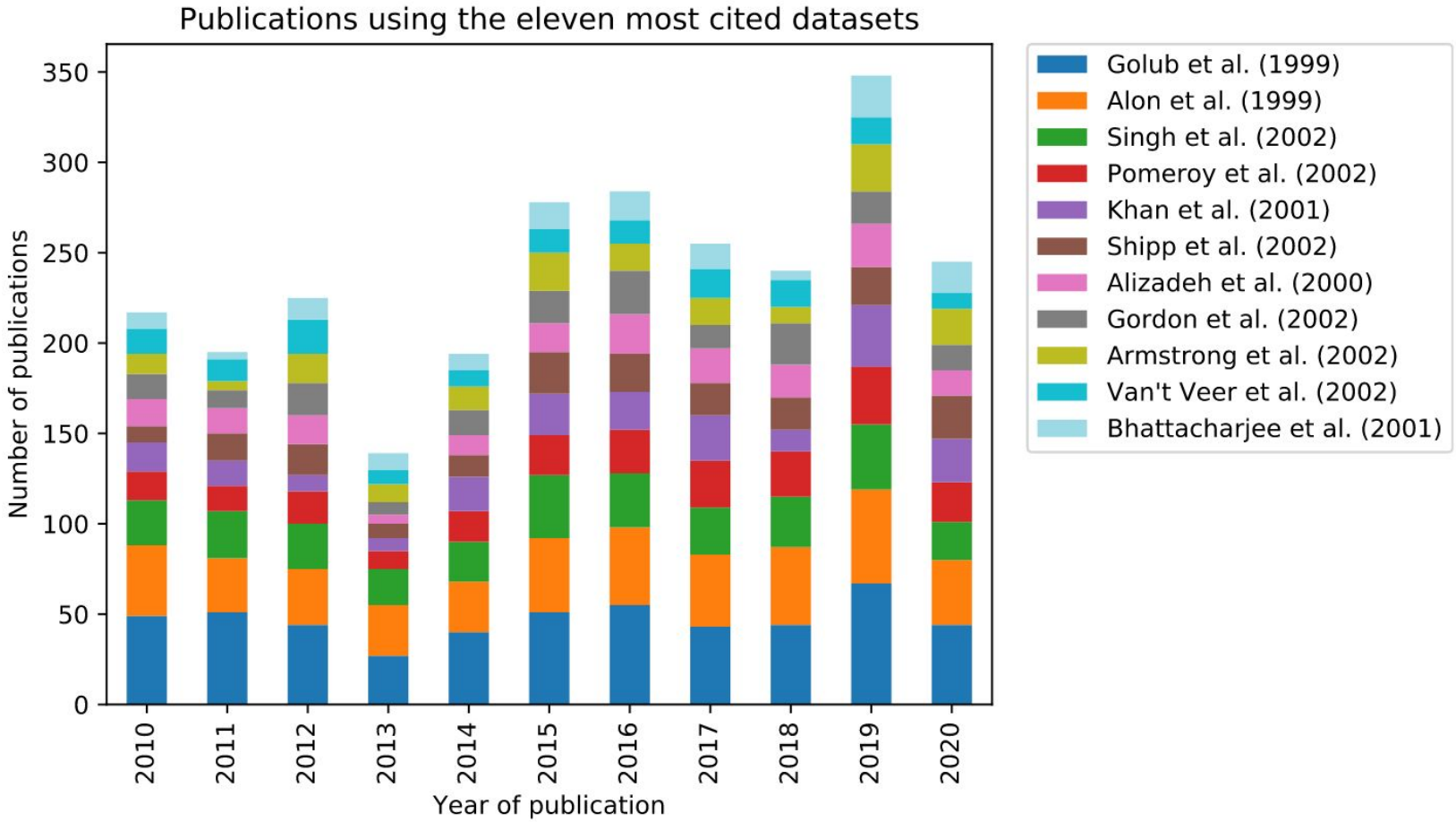
If you entered the right address, you can:

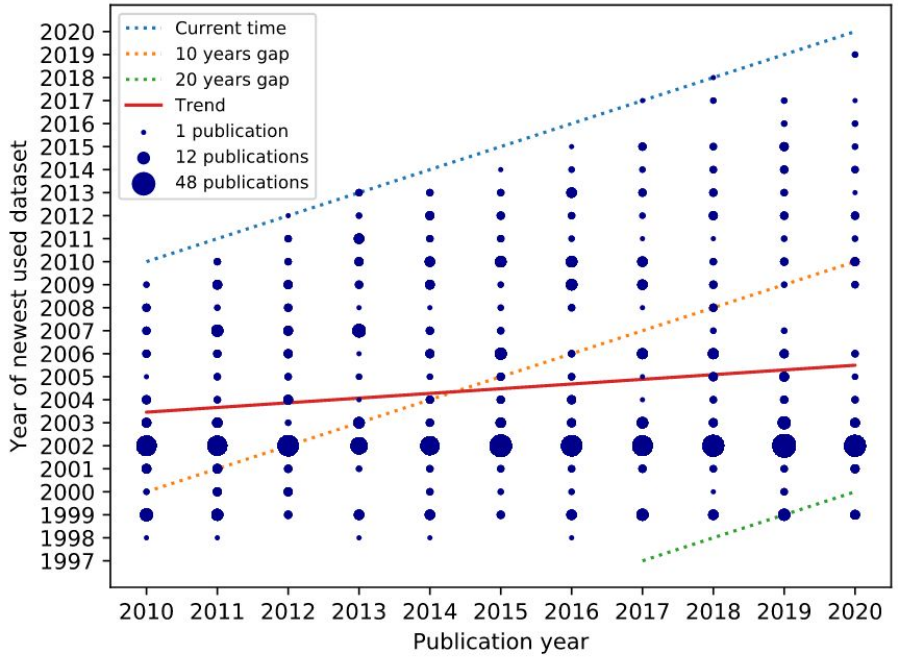
- Try again later
- Check your network connection
- Check that Firefox has permission to access the web (you might be connected but behind a firewall)

Try Again

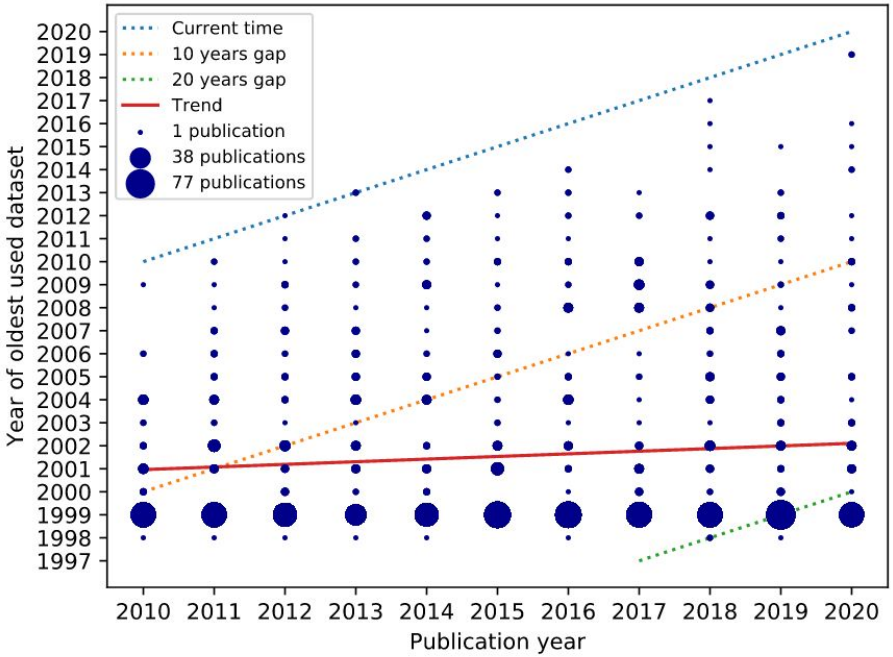
Ranking	Dataset	Year	Samples	Features	Classes	Background	Prevalence in the reviewed publications (%)
1	Golub et al. (1999)	1999	72	7129	2	Leukemia	40.26
2	Alon et al. (1999)	1999	62	2000	2	Colon cancer	32.13
3	Singh et al. (2002)	2002	136	12,600	2	Prostate cancer	22.98
4	Pomeroy et al. (2002)	2002	90	5920	5	Brain cancer	17.67
5	Khan et al. (2001)	2001	83	2309	4	SRBCT	15.94
6	Shipp et al. (2002)	2002	77	7129	2	DLBCL	14.54
7	Alizadeh et al. (2000)	2000	96	4026	9	Lymphoma	13.60
8	Gordon et al. (2002)	2002	181	12,533	2	Lung cancer	13.52
9	Armstrong et al. (2002)	2002	72	11,225	3	Leukemia	12.58
10	Van't Veer et al. (2002)	2002	97	24,481	2	Breast cancer	11.18
11	Bhattacharjee et al. (2001)	2001	203	12,601	5	Lung cancer	10.39

Abbreviations: DLBCL, diffuse large B-cell lymphoma; SRBCT, small round blue cell tumor.

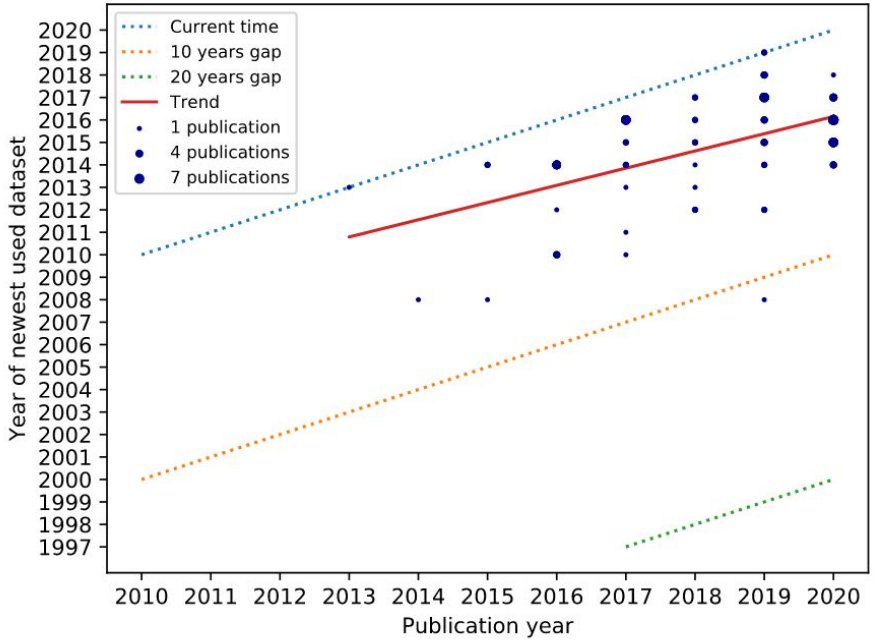

Publications using the eleven most cited datasets
Number of publications
0
50
100
150
200
250
300
350
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year of publication
Golub et al. (1999)
Alon et al. (1999)
Singh et al. (2002)
Pomeroy et al. (2002)
Khan et al. (2001)
Shipp et al. (2002)
Alizadeh et al. (2000)
Gordon et al. (2002)
Armstrong et al. (2002)
Van't Veer et al. (2002)
Bhattacharjee et al. (2001)



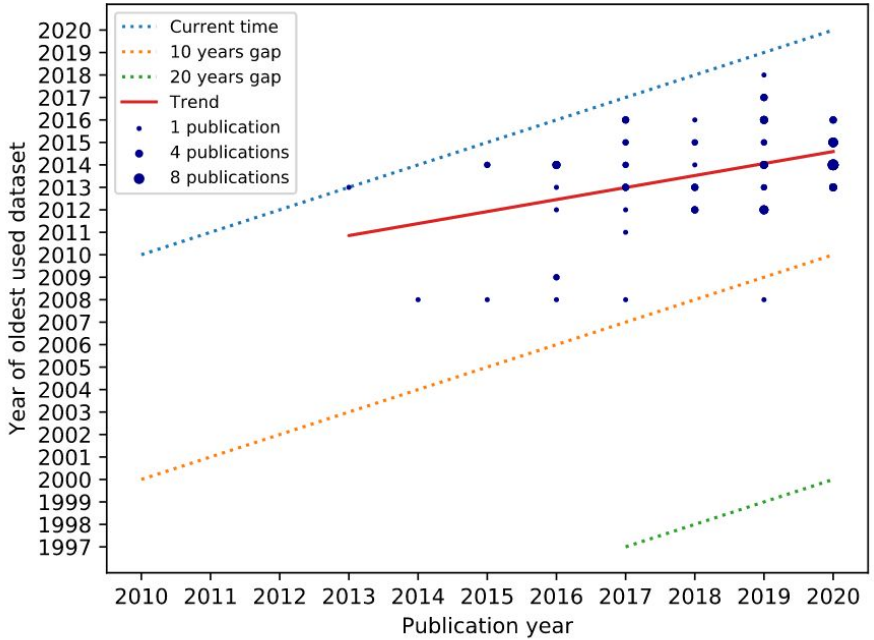
(a) Most recent microarray datasets



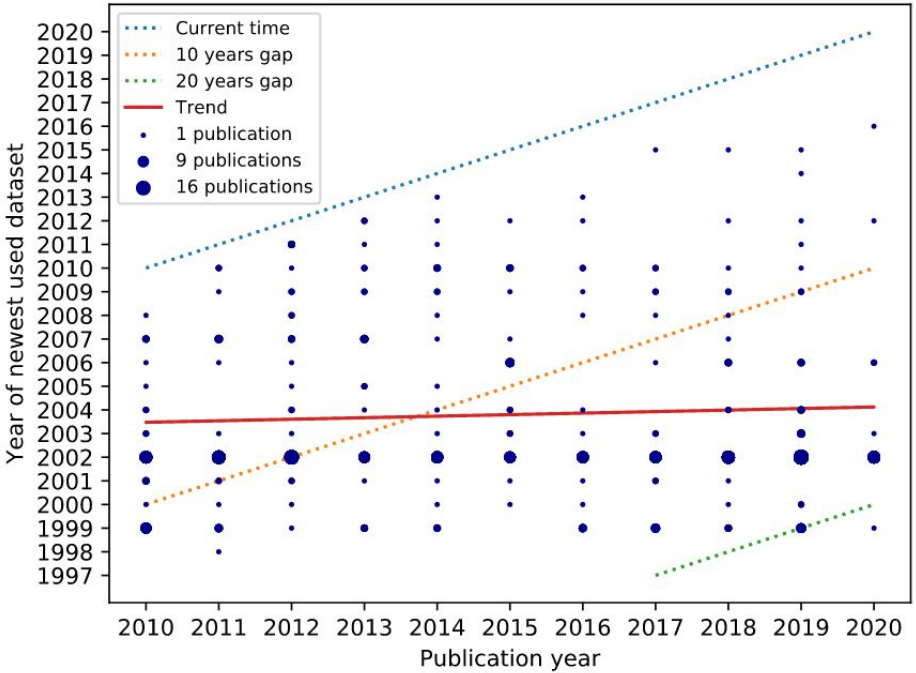
(b) Least recent microarray datasets



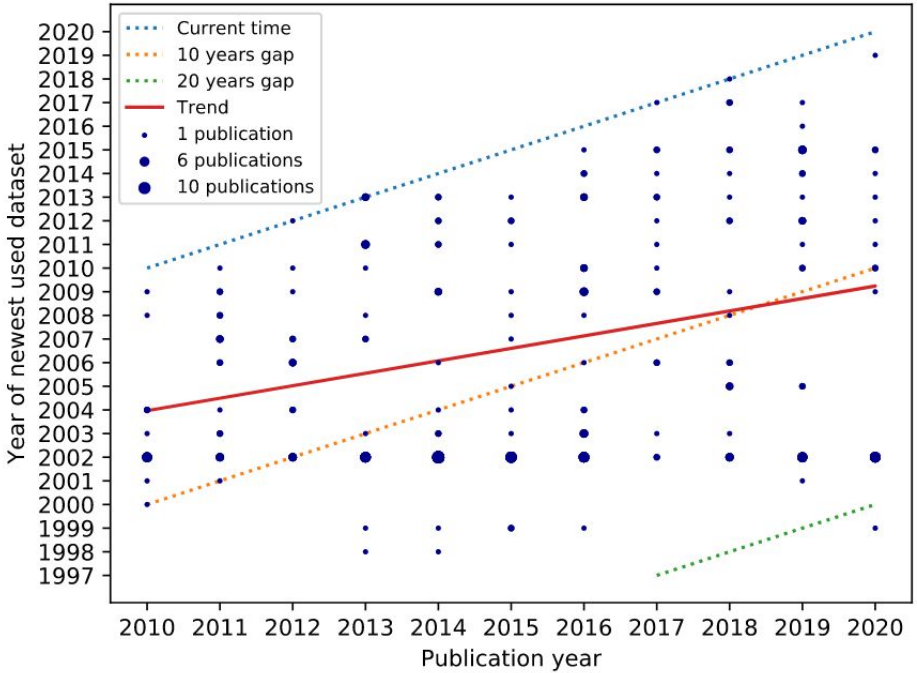
(c) Most recent RNA-seq datasets



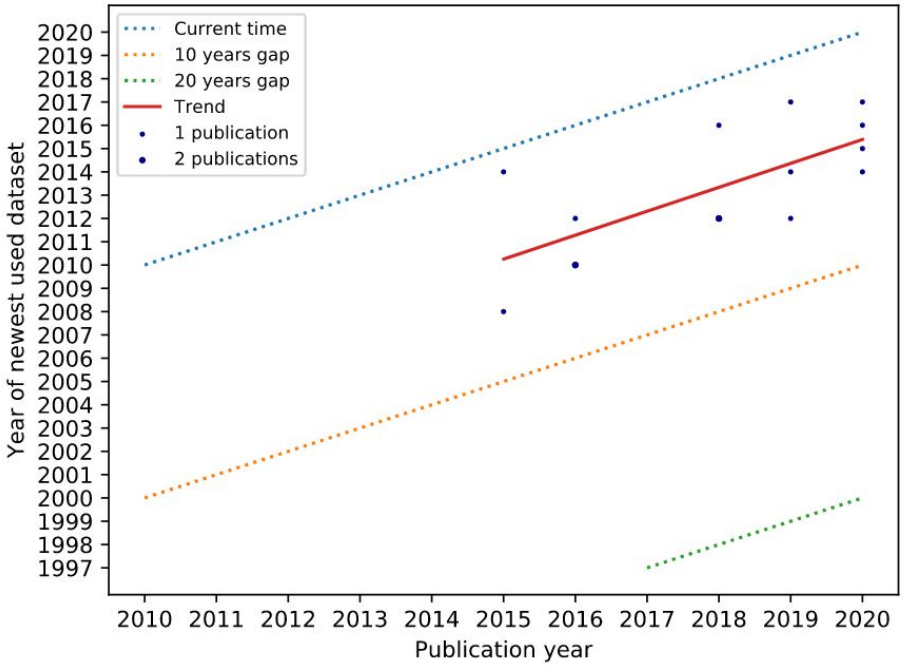
(d) Least recent RNA-seq datasets



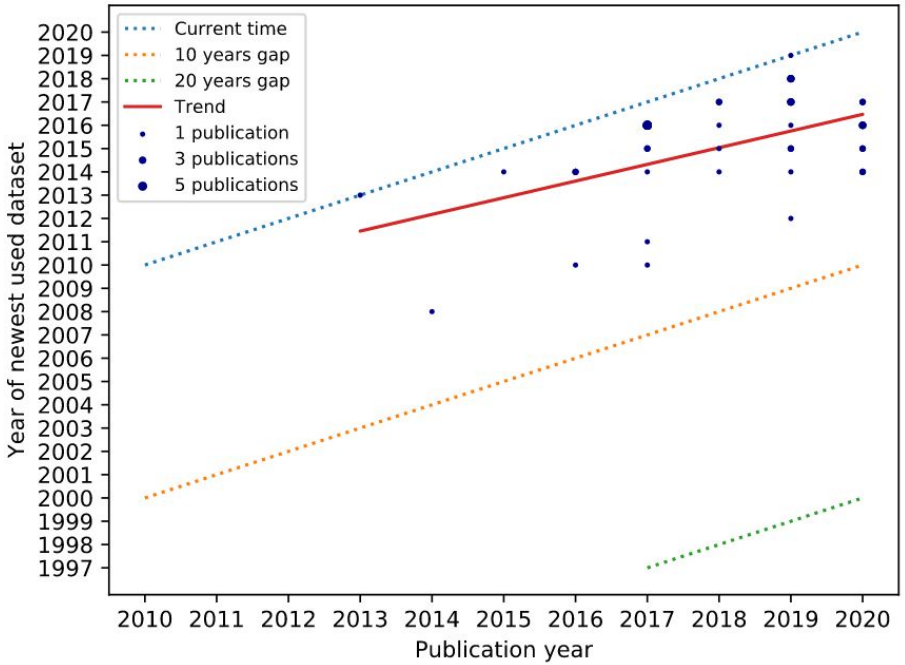
(a) IEEE Xplore — Microarray



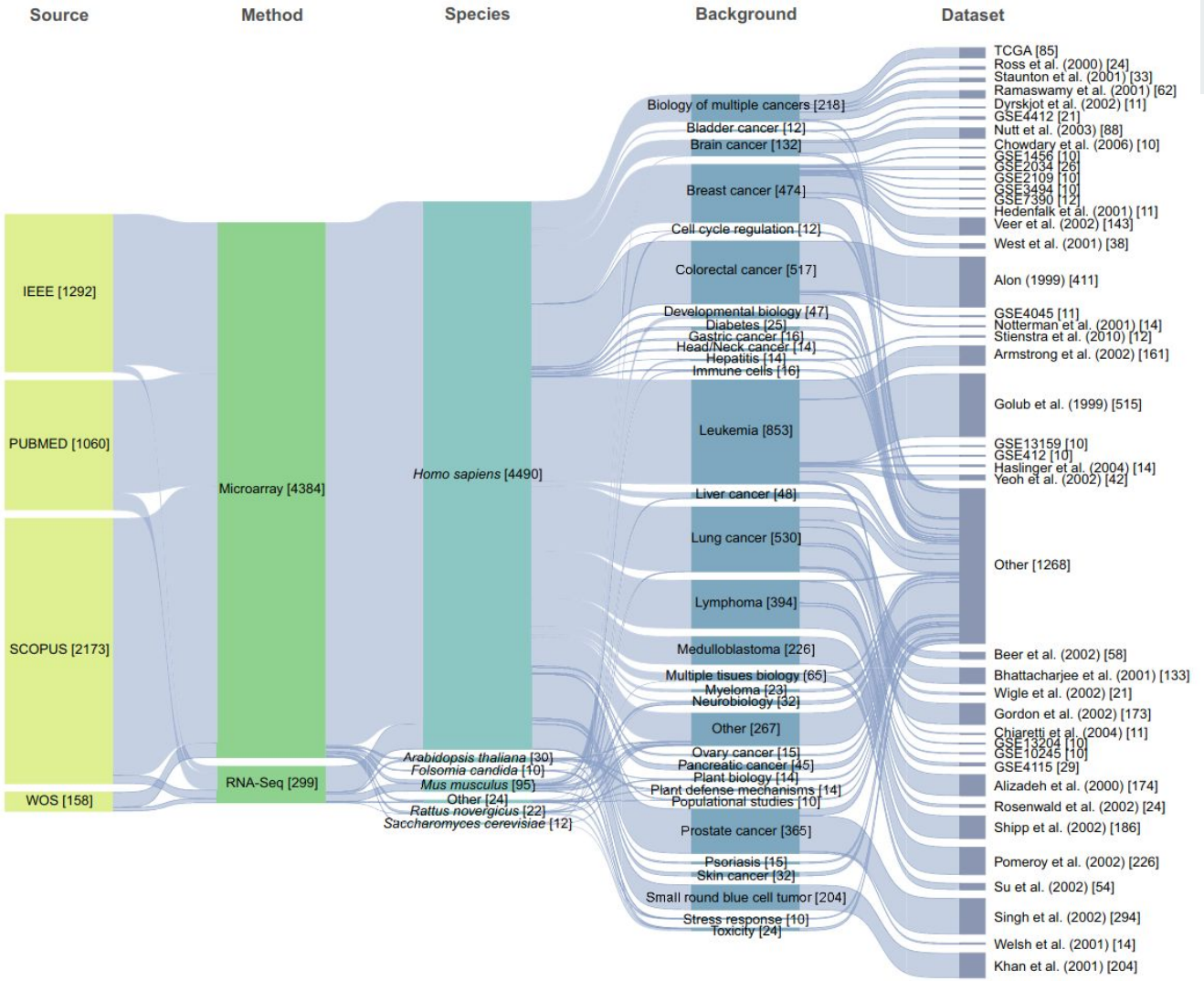
(b) PubMed — Microarray



(c) IEEE Xplore — RNA-seq

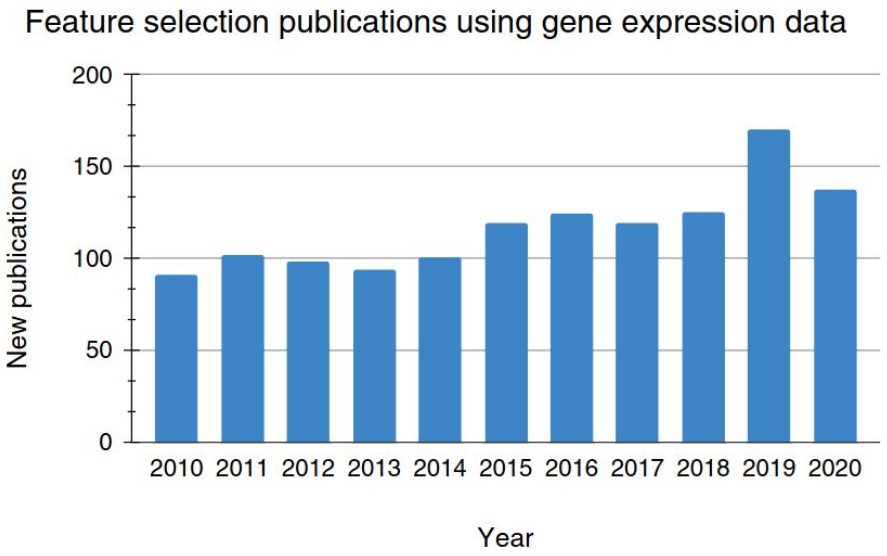


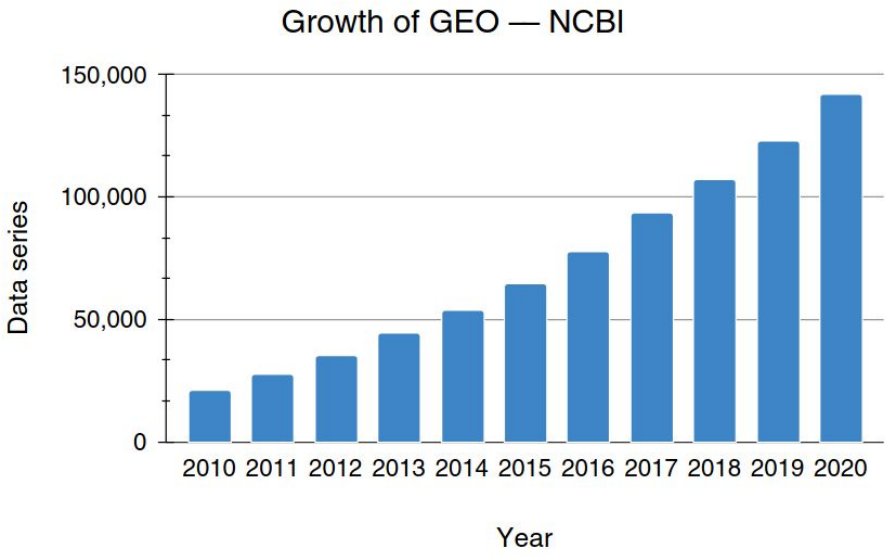
(d) PubMed — RNA-seq

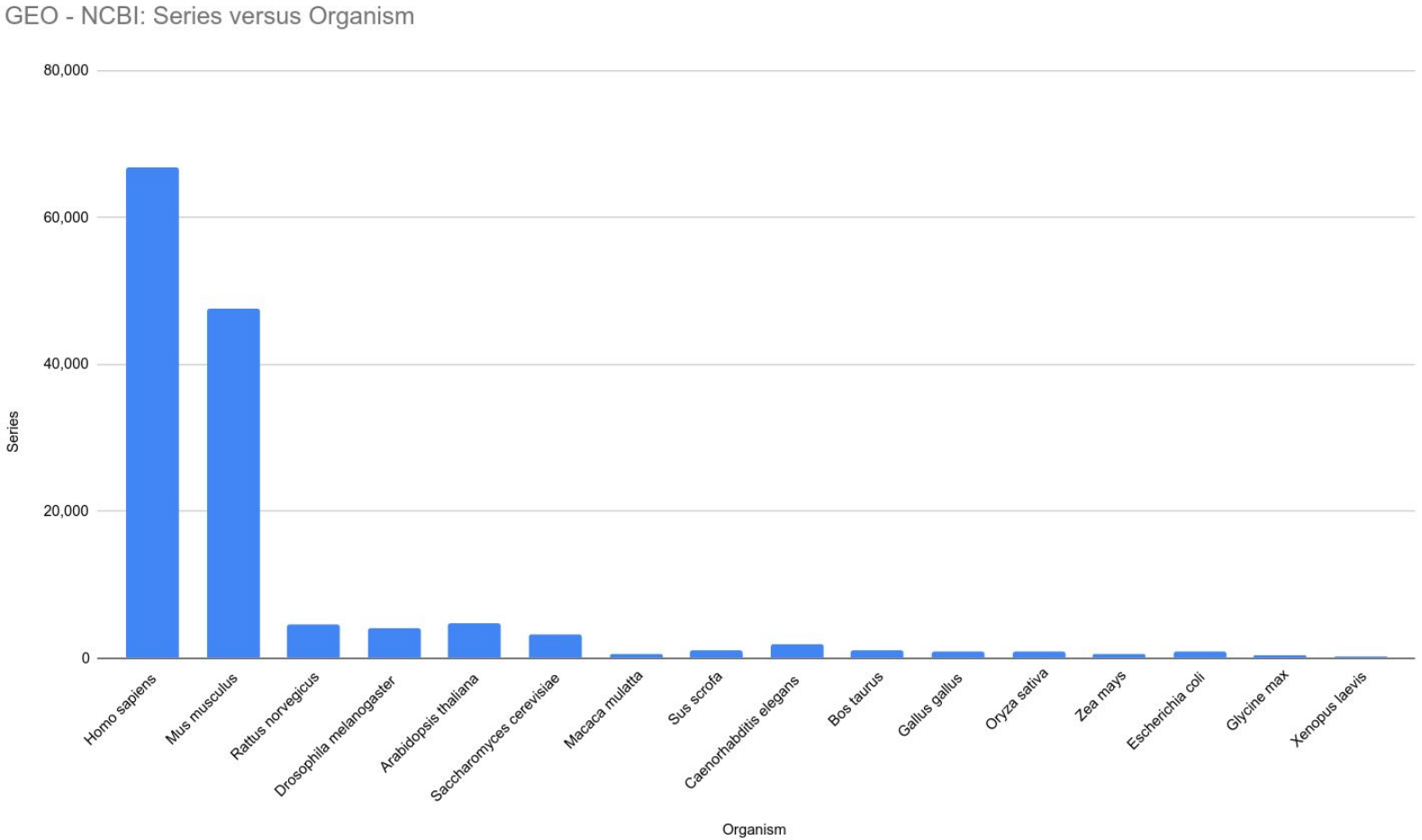

Source
Method
Species
Background
Dataset
IEEE [1292]
PUBMED [1060]
SCOPUS [2173]
WOS [158]
Microarray [4384]
RNA-Seq [299]
Homo sapiens [4490]
Arabidopsis thaliana [30]
Folsomia candida [10]
Mus musculus [95]
Other [24]
Rattus novergicus [22]
Saccharomyces cerevisiae [12]
Biology of multiple cancers [218]
Bladder cancer [12]
Brain cancer [132]
Breast cancer [474]
Cell cycle regulation [12]
Colorectal cancer [517]
Developmental biology [47]
Diabetes [25]
Gastric cancer [16]
Head/Neck cancer [14]
Hepatitis [14]
Immune cells [16]
Leukemia [853]
Liver cancer [48]
Lung cancer [530]
Lymphoma [394]
Medulloblastoma [226]
Multiple tisues biology [65]
Myeloma [23]
Neurobiology [32]
Other [267]
Ovary cancer [15]
Pancreatic cancer [45]
Plant biology [14]
Plant defense mechanisms [14]
Populational studies [10]
Prostate cancer [365]
Psoriasis [15]
Skin cancer [32]
Small round blue cell tumor [204]
Stress response [10]
Toxicity [24]
TCGA [85]
Ross et al. (2000) [24]
Staunton et al. (2001) [33]
Ramaswamy et al. (2001) [62]
Dyrskjot et al. (2002) [11]
GSE4412 [21]
Nutt et al. (2003) [88]
Chowdary et al. (2006) [10]
GSE1456 [10]
GSE2034 [26]
GSE2109 [10]
GSE3494 [10]
GSE7390 [12]
Hedenfalk et al. (2001) [11]
Veer et al. (2002) [143]
West et al. (2001) [38]
Alon (1999) [411]
GSE4045 [11]
Notterman et al. (2001) [14]
Stienstra et al. (2010) [12]
Armstrong et al. (2002) [161]
Golub et al. (1999) [515]
GSE13159 [10]
GSE412 [10]
Haslinger et al. (2004) [14]
Yeoh et al. (2002) [42]
Other [1268]
Beer et al. (2002) [58]
Bhattacharjee et al. (2001) [133]
Wigle et al. (2002) [21]
Gordon et al. (2002) [173]
Chiaretti et al. (2004) [11]
GSE13204 [10]
GSE10245 [10]
GSE4115 [29]
Alizadeh et al. (2000) [174]
Rosenwald et al. (2002) [24]
Shipp et al. (2002) [186]
Pomeroy et al. (2002) [226]
Su et al. (2002) [54]
Singh et al. (2002) [294]
Welsh et al. (2001) [14]
Khan et al. (2001) [204]

Tem motivo?

Tem motivo?


Feature selection publications using gene expression data
New publications
200
150
100
50
0
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
Year


Growth of GEO — NCBI
Data series
150,000
100,000
50,000
0
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
Year


GEO - NCBI: Series versus Organism
80,000
60,000
40,000
20,000
0
Series
Homo sapiens
Mus musculus
Rattus norvegicus
Drosophila melanogaster
Arabidopsis thaliana
Saccharomyces cerevisiae
Macaca mulatta
Sus scrofa
Caenorhabditis elegans
Bos taurus
Gallus gallus
Oryza sativa
Zea mays
Escherichia coli
Glycine max
Xenopus laevis
Organism

E qual o problema?

E qual o problema?

Singh et al. (2002): A distribuição de classes no conjunto de treinamento (49%/51%) é diferente do conjunto de teste (26%/74%). Além disso, este conjunto de teste tem uma diferença quase 10 vezes maior na intensidade do *microarray* em comparação com o conjunto de treinamento.

Gordon et al. (2002): divisão de 50%/50% no treinamento, mas de 90%/10% no teste. Um único gene presente nos dados de treinamento é capaz de classificar corretamente todas as amostras. No entanto, esse mesmo gene não é relevante no conjunto de teste, que, por sua vez, não é linearmente separável.

Bhattacharjee et al. (2001): Mramor et al. (2007) identificaram sete amostras rotuladas incorretamente.

E qual o problema?

Golub et al. (1999): 6817 genes e 38 amostras, sendo pequeno mesmo em comparação com dados dos anos 2000.

Altamente heterogêneo, incluindo amostras de sangue periférico, não apenas de medula óssea, além de pacientes infantis e laboratórios com diferentes protocolos de preparação de amostras.

Usando o método "análise de vizinhança" os autores relataram 100% de precisão na predição de classes. Esses resultados foram altamente insensíveis à seleção particular de genes, com preditores usando de 10 a 200 genes diferentes, todos alcançando a mesma precisão de 100%.

E qual o problema?

Alon et al. (1999) e Khan et al. (2001): experimentos de microarranjo com 6567 genes.

Os autores filtraram o número de genes para 2000 e 2308.

As versões desses conjuntos de dados empregadas por outros pesquisadores foram suas variações filtradas, não os dados brutos originais

Usar dados pré-filtrados para avaliar os selecionadores de atributos pode enviesar os resultados.

Os autores podem estar inadvertidamente misturando os resultados de um algoritmo de filtro de terceiros com os seus próprios.

E qual o problema?

Pomeroy et al. (2002) e **Shipp et al. (2002)** usaram uma plataforma de microarranjo com apenas 6817 sondas, anterior ao fim do HGP e três vezes menor do que o frequentemente usado Affymetrix U133 GeneChip.

Os links para os resultados originais de **Pomeroy et al. (2002)** e o protocolo detalhado de análise de expressão gênica de **Armstrong et al. (2002)** foram armazenados em sites privados que não estão mais funcionais, um problema comum com conjuntos de dados publicados antes do GEO.

Van't Veer et al. (2002): informação das sequências de cRNA para microarranjos, que foram extraídas antes do fim do HGP, contendo, assim, vários vieses pelos padrões atuais.

Possíveis causas?

Possíveis causas

Falta de familiaridade com os dados

Comparação com resultados anteriores

Facilidade de acesso

Bases de dados

Ao longo dos anos, grandes bancos de dados de *machine learning* se tornaram disponíveis.

Esses bancos de dados contêm centenas a milhares de conjuntos de dados de vários domínios.

Objetivo principal dessas plataformas: facilitar a distribuição e acesso de conjuntos de dados.

Foco em experimentos e *benchmarks* de *machine learning*.

Podem amplificar certos vieses ao promover conjuntos de dados já populares.

Gene expression dataset (Golub et al.)

Data | Code (83) | Discussion (4) | Metadata

246 | New Notebook | Download (1 MiB)

Activity Overview

ACTIVITY STATS

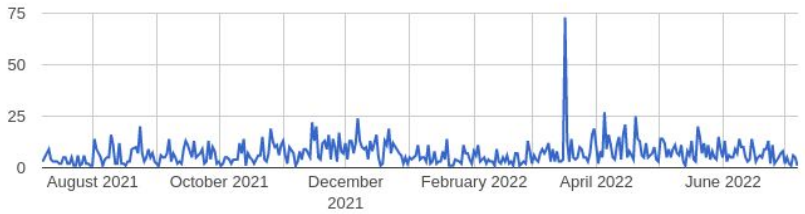
VIEWS
104077

DOWNLOADS
12399

DOWNLOAD PER VIEW RATIO
0.12

TOTAL UNIQUE CONTRIBUTORS
77

Downloads



NOTEBOOKS STATS

NOTEBOOKS
83

NOTEBOOK COMMENTS
104

UPVOTE PER NOTEBOOK RATIO
3.96

NOTEBOOK UPVOTES
329

TOP CONTRIBUTORS

[redacted]

DISCUSSION STATS

TOPICS
4

TOTAL COMMENTS
9

UPVOTE PER POST RATIO
2.56

DISCUSSION UPVOTES
23

Bases de dados

Outros bancos de dados contêm dados biológicos específicos.

Voltados principalmente para biologia.

Organização, jargão e formatos de arquivo podem afastar usuários de outras áreas.

Alguns bancos de dados oferecem conjuntos de dados de expressão gênica curados para uso em pesquisas de aprendizado de máquina.

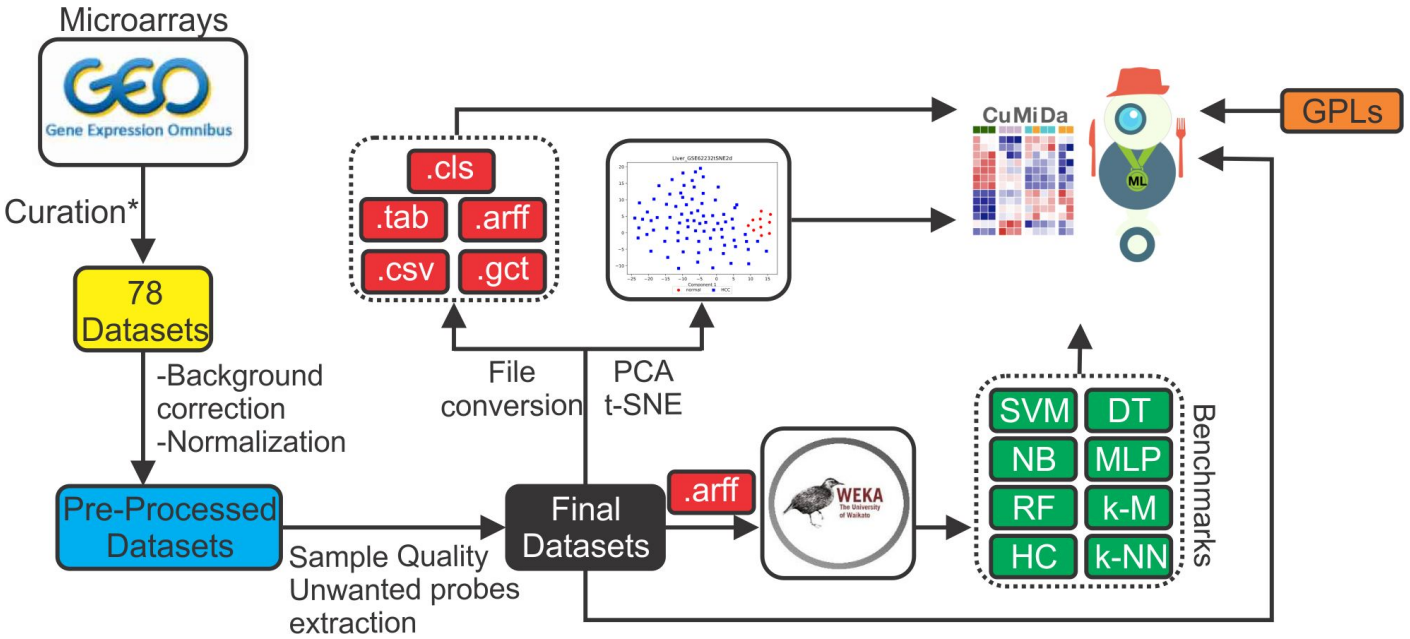
TABLE 3 The most famous public microarray data repositories according to the review by Bolón-Canedo et al. (2014).

Repositories	URL	Status
ArrayExpress European Bioinformatics Institute	http://www.ebi.ac.uk/arrayexpress/	✓
Cancer Program Data Sets Broad Institute	http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi	✖
Dataset Repository Bioinformatics Research Group of Universidad Pablo de Olavide	http://www.upo.es/eps/bigs/datasets.html	✖
Feature Selection Datasets Arizona State University	http://featureselection.asu.edu/datasets.php	✖
Gene Expression Model Selector Vanderbilt University	http://www.gems-system.org	✖
Gene Expression Omnibus National Institutes of Health	http://www.ncbi.nlm.nih.gov/geo/	✓
Gene Expression Project Princeton University	http://genomics-pubs.princeton.edu/oncology/	✓
Kent Ridge Bio-Medical Dataset Repository Agency for Science, Technology and Research	http://datam.i2r.a-star.edu.sg/datasets/krbd	✖
Stanford Microarray Database Stanford University	http://smd.stanford.edu/	✖

Note: The sources and the provided URLs are listed as informed by Bolón-Canedo et al. (2014), which accessed all repositories in January 2014. In less than a decade, most URLs are not working anymore (marked with 55). The current status of the URLs was checked on October 14, 2023.

Databases	Curated	Source	Quality control[a]	Up to date[b]	File formats[c]
ARCH4	No	Normalized; gene and transcript level counts	No	Yes	.h5
BARRA:CuRDa	Yes	Normalized	Yes	Yes	.csv; .tab; .gct; .cls
BioLab	No	Author's	No	No	.tab
CuMiDa	Yes	Normalized	Yes	Yes	.csv; .tab; .gct; .cls; .arff
Datamicroarray	No	Author's	No	No	.r; .RData
Gene Expression Project	No	Author's	No	No	.tab; .xls
InSilicodb	Yes	Varies	NS	Yes	.r
PSO-EMT datasets	No	Varies	No	No	.mat
Recount3	Yes	Gene and transcript level counts	No	Yes	.txt; .bw; .mtx; .RData
Refine.bio	No	Normalized; gene and transcript level counts	No	Yes	.sf; .json; .tsv
RNASeq-er	No	Normalized; gene and transcript level counts	No	Yes	.cram; .bw; .bedGraph

CuMiDa


Microarrays
Gene Expression Omnibus
Curation*
78 Datasets
-Background correction
-Normalization
Pre-Processed Datasets
Sample Quality
Unwanted probes
extraction
.cls
.tab
.arff
.csv
.gct
File conversion
PCA
t-SNE
Final Datasets
.arff
WEKA
The University of Waikato
CuMiDa
ML
GPLs
SVM
DT
NB
MLP
RF
k-M
HC
k-NN
Benchmarks

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 26, Number 0, 2019
© Mary Ann Liebert, Inc.
Pp. 1–11
DOI: 10.1089/cmb.2018.0238

Research Article

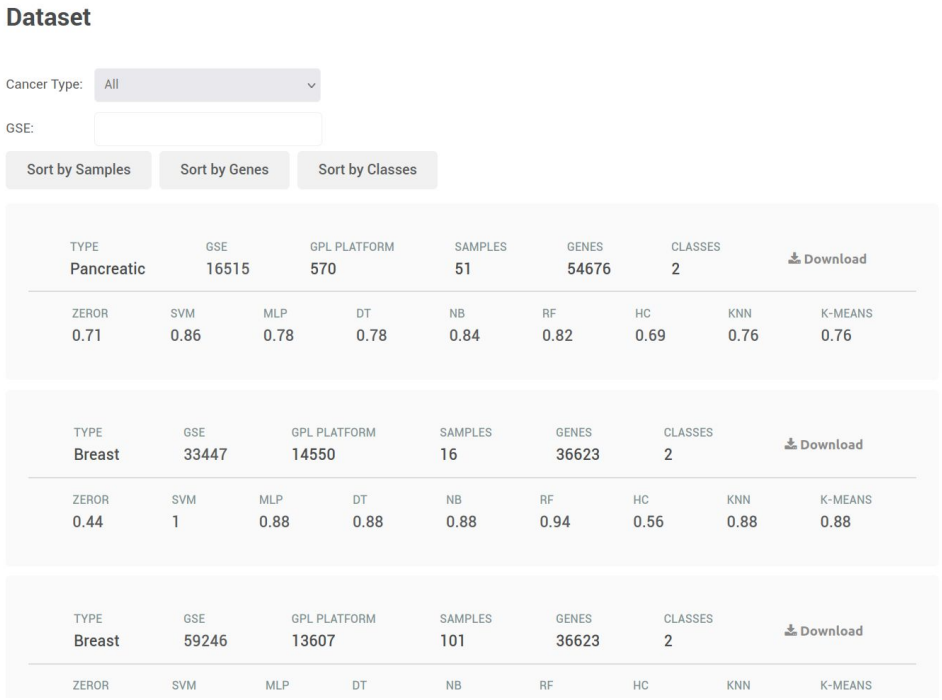
CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research

BRUNO CÉSAR FELTES, EDUARDO BASSANI CHANDELIER, BRUNO IOCHINS GRISCI, and MÁRCIO DORN

ABSTRACT

The employment of machine learning (ML) approaches to extract gene expression information from microarray studies has increased in the past years, specially on cancer-related works. However, despite this continuous interest in applying ML in cancer biomedical research, there are no curated repositories focused only on providing quality data sets exclusively for benchmarking and testing of such techniques for cancer research. Thus, in this work, we present the *Curated Microarray Database* (CuMiDa), a database composed of 78 handpicked microarray data sets for *Homo sapiens* that were carefully examined from more than 30,000 microarray experiments from the *Gene Expression Omnibus* using a rigorous filtering criteria. All data sets were individually submitted to background correction, normalization, sample quality analysis and were manually edited to eliminate erroneous probes. All data sets were tested using principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) analyses to observe sample division and were additionally tested using various ML approaches to provide a base accuracy for the major techniques employed for microarray data sets. CuMiDa is a database created solely for benchmarking and testing of ML approaches applied to cancer research.

Keywords: benchmarking, cancer, classification, curation, machine learning, microarray, supervised learning, unsupervised learning.



https://doi.org/10.1089/cmb.2018.0238

https://sbcb.inf.ufrgs.br/cumida

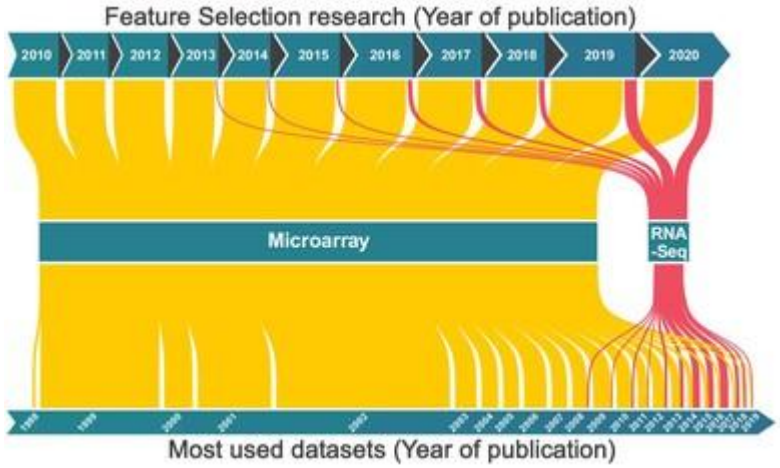
CuMiDa

TABLE 3. MAJOR CRITERIA DESIGNED TO EVALUATE A PROPER BENCHMARK

Criteria	*CuMiDa*
Active benchmarks in research area	Results obtained from the top, most used ML tools in the field
Trustworthy evaluated tools	All tools available at WEKA software
Transparency with conducted protocols	Full description: from data mining filtering criteria to classification protocol
Availability of benchmarked in/outputs	All inputs and outputs fully available at the database
Relevance of employed metrics	All metrics are the most used in the field of cancer microarray analysis
Availability of benchmarked methods	All methods can be found in the WEKA software
No inclusion of new tools	All benchmark results are from traditional methods

ML, machine learning.

Artigo final



Received: 2 February 2023 | Revised: 27 October 2023 | Accepted: 30 October 2023
DOI: 10.1002/widm.1523

ADVANCED REVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

The use of gene expression datasets in feature selection research: 20 years of inherent bias?

Bruno I. Grisci[1,2] | Bruno César Feltes[3] | Joice de Faria Poloni[3] | Pedro H. Narloch[1] | Márcio Dorn[1,4,5]

[1]Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil
[2]Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada
[3]Institute of Biosciences, Federal University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil
[4]National Institute of Science and Technology - Forensic Science, Porto Alegre, Rio Grande do Sul, Brazil
[5]Center for Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

Correspondence
Márcio Dorn, Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, 91501-970, Brazil.
Email: mdorn@inf.ufrgs.br

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Numbers: 151591/2022-9, 314082/2021-2, 408154/2022-5, 440279/2022-4, 465450/2014-8; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Finance Code 001, Grant/Award Numbers: 88881.198766/2018-01, 88881.522073/2020-01; Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul, Grant/Award Numbers: 17/2551-0000520-1, 19/2551-0001906-8; Global Affairs Canada, Emerging Leaders in the Americas Program Scholarship

Abstract
Feature selection algorithms are frequently employed in preprocessing machine learning pipelines applied to biological data to identify relevant features. The use of feature selection in gene expression studies began at the end of the 1990s with the analysis of human cancer microarray datasets. Since then, gene expression technology has been perfected, the Human Genome Project has been completed, new microarray platforms have been created and discontinued, and RNA-seq has gradually replaced microarrays. However, most feature selection methods in the last two decades were designed, evaluated, and validated on the same datasets from the microarray technology's infancy. In this review of over 1200 publications regarding feature selection and gene expression, published between 2010 and 2020, we found that 57% of the publications used at least one outdated dataset, 23% used only outdated data, and 32% did not cite data sources. Other issues include referencing databases that are no longer available, the slow adoption of RNA-seq datasets, and bias toward human cancer data, even for methods designed for a broader scope. In the most popular datasets, some being 23 years old, mislabeled samples, experimental biases, distribution shifts, and the absence of classification challenges are common. These problems are more predominant in publications with computer science backgrounds compared to publications from biology and can lead to inaccurate and misleading biological results.

This article is categorized under:
Algorithmic Development > Biological Data Mining
Technologies > Machine Learning

KEYWORDS
feature selection, gene expression, machine learning, microarray, RNA-seq

https://doi.org/10.1002/widm.1523

Problema resolvido?

Recomendações

- Citar a fonte original dos dados.
- Mencionar as principais características dos conjuntos de dados, como o número de amostras, atributos e classes.
- Se vários conjuntos de dados forem usados, listar de forma organizada.
- Adicionar *hyperlinks* funcionais para a fonte dos dados. Se os dados estiverem em bancos de dados privados, verificar se não estão disponíveis em repositórios públicos.
- Usar vários conjuntos de dados com propriedades distintas para validação.

Recomendações

- Não usar apenas dados de *H. sapiens* e de câncer.
- Não usar conjuntos de dados pré-filtrados pelo autor.
- Evitar usar conjuntos de dados antigos.
- Usar dados de RNA-seq.
- Ao projetar um novo banco de dados de expressão gênica para aprendizado de máquina, seguir as diretrizes propostas por **Peters et al. (2018), Wilkinson et al. (2016), Walsh et al. (2021)** e **Hutchinson et al. (2021)**.
- Revisores e editores devem garantir a aplicação dos itens acima nas publicações.
- Usar métodos interpretáveis de aprendizado de máquina e visualização para melhorar a compreensão e a replicabilidade dos resultados de experimentos.