3I4-3

Wildcard 許容頻出部分グラフパターンのグラフ分類への応用

Applications of Frequent Subgraphs with Wildcards to Graph Classification

岡崎　文哉 *1　　瀧川　一学 *2*3
Fumiya Okazaki　　Ichigaku Takigawa

*1北海道大学工学部
School of Engineering, Hokkaido University

*2北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido Universit

*3科学技術振興機構，さきがけ
JST,PRESTO

In this paper, we empirically study the use of frequent subgraphs with wildcards for graph classification. Incorporating wildcards into subgraph feature representations enables to provide more flexible indicators by smaller subgraph patterns, which are obtained in a shallower search. Our experimental evaluations show the improvement on the classification accuracy for the majority of 13 used datasets. We also investigate and discuss the parameter dependency and comparisons to subgraph patterns without wildcards.

1. はじめに

グラフは，低分子化合物の構造式 [Takigawa 13]，RNA 二次構造 [Karklin 05]，自然言語処理 [Kudo 05] などの知識処理に幅広く用いられている重要なデータ構造である．近年こうした科学分野のグラフデータが PubChem など公共のデータベースに蓄積・整備されるようになり，有効な利活用が喫緊の課題となっている．特に，グラフデータからの教師付き学習は，生命科学や物質科学における構造活性相関や構造物性相関の定量的モデルとして機械学習分野で研究されており，より高精度で高効率的な手法が求められている [瀧川 16].

本稿では，グラフ分類として，グラフデータからの教師付き学習を扱う．グラフデータからの教師付き学習では，特徴量として部分グラフの有無を用いることが多く，この部分グラフの特徴選択が学習モデルの精度を左右する．しかし，与えられたグラフデータに起こるすべての部分グラフを列挙することは現実的に不可能である．したがって，グラフカーネル法 [Kashima 03] や ECFP 法 [Rogers 10] など，対象となる部分グラフのクラスを予め発見的に制限する手法や，gBoost 法 [Saigo 09] や Adaboost に基づく手法 [Kudo 05] など，必要な部分グラフのみを効率的に探索しながら学習を行う手法が提案されている．全ての部分グラフの列挙は現実的に不可能であるため，特徴として使う頻出部分グラフは，閾値 σ(与えられたグラフデータベースに対して σ 以上のグラフに出現する) を与え，列挙することを考える． その時，σ を小さくしたほうが特徴候補の数を増やすことができるため精度向上が期待できる．しかし，σ を下げると出力数が指数的に増加するため，深い探索が必要な事例では有効な特徴がある場合でもメモリや時間に対するトレードオフがある．

本稿では，我々が提案した wildcard を許容した頻出部分グラフマイニング [岡崎 15] のグラフ分類への応用を提案する．グラフ分類において使用する部分グラフ特徴は，包含関係に由来する冗長性や相関を持っており，wildcard を許容した緩和特徴表現により，厳密表現と比べて小さい部分グラフで柔軟な記述力を得ることが期待できる．そして，wildcard 許容頻出部分グラフを特徴に用いた場合の効果を確かめるため，13 種類の低分子化合物の構造活性相関データを用いて，実験的に検証する．また，通常の頻出部分グラフを特徴に用いた場合やパラメタを変化させた場合と比較し，wildcard 許容頻出部分グラフの有効性と条件の分析を行う．

第 2 節では，頻出部分グラフを用いたグラフ分類，第 3 節では，グラフ分類に用いた機械学習の手法 Random Forest の説明を行う．第 4 節では，wildcard 許容頻出部分グラフによるグラフ分類を提案する．第 5 節では，提案手法に対する実験を示し，第 6 節で，実験に対する考察を行う．第 7 節で，結論を述べる．

連絡先: 岡崎　文哉，北海道大学工学部情報エレクトロニクス学科，fokazaki@art.ist.hokudai.ac.jp

2. 頻出部分グラフによるグラフ分類

本稿において，グラフ分類とは，グラフ集合に対する教師付き学習を指す．複数の種類のグラフ集合が与えられた時，これらのグラフ集合を分けるための特徴付けや与えられたグラフがいずれの集合に属するかを判別する問題である．

本稿では，2 種類のグラフ集合を入力として与え，分類器を構築する．分類器の構築において，特徴量の決定が重要な課題である．この特徴量の決定に対して部分パス，部分木，部分グラフなど様々な特徴量を考えることができるが，我々は頻出部分グラフによる特徴付けに着目した．

すべての部分グラフパターンの列挙は，出力数が組み合わせ爆発を起こすため，現実的に求めることは困難である．そのため，通常，ある閾値を与え，σ 個以上のグラフに出現するパターンを頻出部分グラフとして列挙する．この閾値 σ を最小支持度（以下，minsup）と呼ぶ．

頻出部分グラフによるグラフ分類では，minsup によって精度が異なる．原理上は，minsup を下げることで，得られる部分グラフパターンの集合は単調に増大するので，用いる機械学習アルゴリズムが適切に対応できれば，精度も単調に向上する．しかし，minsup を下げれば出力数 (用いる特徴の数) は指数的に増加するため、計算時間や使用メモリの点で現実的には全列挙が困難となる．そのため，計算可能な特徴量の数には制限があり，この制限に対して分類の精度を上げる必要がある．このような場合，従来は用いるグラフの型をパスや木のみに制限する方法が検討されてきた．本稿では型はグラフのまま変え

ず，wildcard を許容した特徴表現を用いた場合の有効性について実験的に検証を行う．

3. Random Forest

本稿では、グラフ分類で用いる特徴集合の良さの評価が目的であるため，分類に用いる学習アルゴリズムには統一して Random Forest を用いた．

Random Forest[Breiman 01] とは，複数の決定木を統合してマルチクラス分類を行うアンサンブル学習アルゴリズムである．このアルゴリズムの特徴は，ブートストラップを取り入れ過学習を防ぐことが可能な点，Random Feature Selection[Ho 98] を取り入れることで特徴ベクトルの次元数が大きくても高速に学習が可能である点にある．これらのメリットを持ち，クラス識別，クラスタリング，回帰分析を行うことができる．このように高い汎用性を持つため，様々な分野に応用されている．

本稿において，Random Forest を用いた理由は，以上の Random Forest の特性に加えて，各特徴に対して重要度を定めることができ，どの特徴が分類に寄与していたかの説明が定量的にできるためである．

4. 提案手法

本稿では，グラフ分類のための特徴量として，wildcard 許容頻出部分グラフ [岡崎 15] を用いる手法を提案する．本節では，wildcard 許容頻出部分グラフの列挙対象について説明する．

定義 1 (wildcard 許容頻出部分グラフ) *wildcard* を k 個許容した頻出部分グラフは，任意のラベルにマッチするラベルである *wildcard* を k 個以下含む部分グラフで，入力で与えた閾値 *minsup* 以上のグラフに含まれる部分グラフ全てである．

一般的なグラフデータにおいては辺のラベルにおける wildcard が意味を持つことがあると考えられるが，本稿において，取り扱うデータベースは，化合物の分子構造を表したものであり，グラフの辺のラベルは頂点間の結合数を表しているため，この部分に対する wildcard 許容は重要な構造になる場合が少ないと判断することができる．そのため，本稿の列挙対象，その対象実験においては，頂点ラベルに対してのみ wildcard 許容を考える．

頻出部分グラフは minsup を下げることで，列挙対象となるパターン数は膨大に増加する．さらに，wildcard を許容した場合，許容数を増やすことで大幅に列挙数が増加し，適切な対処をしなければ，現実的に列挙できない数になる．我々は，wildcard を許容した頻出部分グラフマイニング [岡崎 15] において，効率的な列挙の方法を提案している．さらに，出力数の増加問題に対して，冗長なパターンとその削減方法，その頻出飽和・極大パターン集合を求める手法を示した．

本稿では，wildcard 許容頻出部分グラフをグラフ分類に対する特徴として使う手法を提案し，wildcard の有効性について検証する．また，単純に wildcard 許容頻出部分グラフを用いる手法に加え，minsup が大きい部分に wildcard を許容し，出力数の増加を抑えるために，通常の頻出部分グラフと組み合わせて用いるという，wildcard 許容の応用を提案し，実験を行う．さらに，冗長なパターンの削減や頻出飽和・極大パターン集合を用いることで，応用するグラフデータベースに対して wildcard 許容頻出部分グラフをうまく利用することができると期待できる．

グラフ分類の特徴量として，wildcard 許容頻出部分グラフを用いる場合，どのパターンが重要なパターンであるかという解釈が困難である．本稿における実験では，wildcard を許容することによる有効性に対して，一般的な知見を得るために，各データベースによる設定の変更は行わず，冗長なパターンの削減もせずに比較を行った．次の節で，詳細について議論する．

CPU	Intel(R) Core(TM) i7-3770K 3.50GHz
メモリ	33GB
OS	Ubuntu 14.04.3 LTS

表 1: 実験環境

データセット名	グラフ数	平均ノード数	平均エッジ数
CPDB	684	14.1	14.6
Mutag	188	17.9	19.8
NCI1	1000	35.3	38.5
NCI41	1000	35.5	38.7
NCI47	1000	35.5	38.7
NCI81	1000	34.4	37.5
NCI83	1000	34.5	37.5
NCI109	1000	35.2	38.4
NCI123	1000	34.2	37.3
NCI145	1000	34.9	38.0
NCI167	1000	30.5	33.5
NCI220	1000	28.3	30.2
NCI330	1000	30.5	33.0

表 2: 使用したデータセット

5. 実験

本節では，wildcard 許容頻出部分グラフをグラフ分類の特徴として使用する実験を示す．本稿では，13 種類のデータベースに対して，通常の頻出部分グラフと wildcard 許容頻出部分グラフを比較することで，wildcard 許容部分グラフの有効性について検証を行う．

5.1 実験環境とデータセット

本実験は表 2 に示す 13 種類のグラフデータを用いて，表 1 に示す環境で行った．データセット名「NCIx」は PubChem BioAssay の AID 番号 x のデータセットである [*1]．これらのデータセットのうちグラフ数が膨大なものは，各データにおける 1000 個のランダムサブセットを用いた．その時，正例と負例の数が同じになるように選択を行った．その際に，正例が 500 に満たないデータに対しては，グラフ数が 1000 になるように負例を追加した．構造活性相関の実験の実データでは通常，負例数が極めて多く，この不均衡性 (imbalance) を考慮した評価尺度を用いる必要がある．例えば，ROC50[Wale 08] を用いるなどが提案されているものの，上位 50 個で良いかなど解釈の問題のため，広く確立した尺度とは言い難い．本稿では単純に正例と負例の数を均衡化し，通常の正解率により特徴集合の良さを評価する．

5.2 実験結果

本実験において，正解率はデータを 10-分割交差検証を行った際のテストデータに対する正解率（ACC と表記）を示す．用いたデータセット１３種類のうち，CPDB，Mutag は minsup を変化させた時の精度の変動を示すために使用した。また，NCIx

*1 http://www.ncbi.nlm.nih.gov/pcassay

のデータを用いて，wildcard を許容した場合の精度と出力数に対する傾向を検証するための実験を行った．

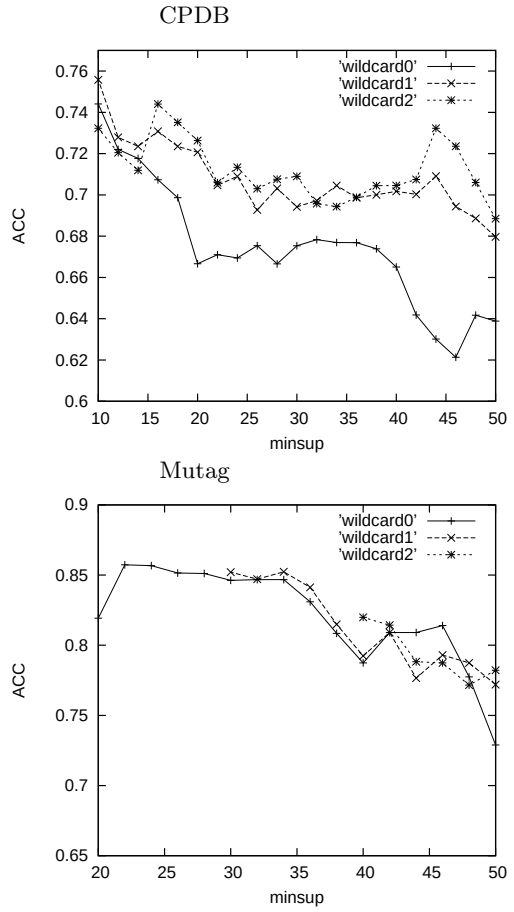

図 1: minsup の変化と正解率の実験結果

図 1 に wildcard 許容に対する minsup を変動させた時の実験結果を示す．上段が CPDB，下段が Mutag に対する実験結果である．この図において,「wildcardx」は wildcard を x 個許容を指す．

図 1 上段の CPDB では，minsup が全体的に精度の向上が見られる．minsup を十分に下げるとあまり精度に差がないが，通常の頻出部分グラフパターンによる結果と同等か，それ以上の結果が出ている．また，wildcard0（wildcard 許容なし）では，minsup が 20%以上になると 0.72-0.74 から 0.66-0.68 に急に精度が下がっている．しかし，wildcard を許容することで minsup が 20%以上でも 0.70-0.72 の精度が得られている．

図 1 下段の Mutag では，minsup が 20%-33%では精度にほぼ変化がなく，33 以上になると，wildcard の許容にかかわらず，0.75-0.8 周辺に精度が落ち込む．wildcard を許容することによる精度の悪化は見られないが，精度の向上も見られないことから，wildcard の有効なデータベースではないことがわかる．

表 3 に，NCI データセットに対する実験結果を示す．この表では，各データセットに対して，通常の頻出部分グラフ，wildcard を 1 つ許容した頻出部分グラフ，wildcard を 2 つ許容した頻出部分グラフについて，minsup を 45%から 5%まで 5%ずつ変化させた場合の出力数と正解率（ACC と表記）を示す．NA と表記した部分は現実験環境でメモリ不足により求めることのできなかったものである．このとき，精度が一番良かったものを太字としている．このデータセットにおいて，11 個中 10 個で wildcard 許容のほうが良い結果が出ている．

さらに，右側には，wildcard 許容と通常の頻出部分グラフを組み合わせた場合の実験結果を示している．これは先ほど太字で示した部分の特徴に加えて，それより minsup が小さい頻出部分グラフを追加した場合の出力数と正解率である．例えば，NCI1 では，wildcard2 許容の 30%で良い結果が出ていた．この時の特徴量 10721 に加えて，通常の頻出部分グラフを minsup10%まで加える．ここで加えている特徴量は 30% 10%における（3283-225）個である．それにより，特徴量は 13778 個であり，精度は wildcard2 許容の 30%に対して向上しているという結果を得た．このように，組み合わせにより，精度が向上したものに対しても太字で示している．実験結果より，wildcard 許容により精度が向上した 10 個のデータセットのうち，3 個のデータセットで，組み合わせによる精度の向上が見られた．

これらの結果に加え，Random Forest による特徴量の重要度の高かったものを見てみたところ，wildcard 許容部分グラフが上位の 20 個程度をほぼ占めていた．これにより，通常の頻出部分グラフより，wildcard 許容部分グラフのほうが分類に寄与する特徴として Random Forest で選ばれやすかったことが分かる．従って，wildcard 許容による拡張で得られた部分グラフ特徴は冗長なものではなく実際に分類に対する有効性を持つものと言える．

6. 考察

実験結果より，wildcard を許容することで，単純に minsup を下げるより，精度が向上するデータベースが多いという結果になった．一般的に，Random Forest は弱学習器として決定木をもつため，正例と負例をよりうまく分ける特徴が多く存在するほど精度が上がる場合が多い．wildcard を許容したパターンは通常のパターンに加え，データベースをうまく表現するような特徴としての候補を得ることができ，このようなパターンが精度の向上に寄与すると考えられる．

また，頻出部分グラフでは，minsup を大きく設定すると，得られるパターン数が少なく表現力が弱い．それに対し，wildcard を許容すると，minsup を大きく設定した場合に表現力が厚みを増すため，精度が向上すると考えられる．しかし，wildcard を許容することによる出力数の増加が原因で minsup を下げて計算することが困難である．そこで，minsup を下げた部分の表現を与えるために頻出部分グラフと組み合わせることで，さらに精度が向上するデータベースが存在することも実験により示すことができた．

7. 結論と今後の課題

本稿では，グラフ分類に対する特徴量として，wildcard 許容頻出部分グラフを利用する手法を提案し，実験により wildcard 許容の有効性について検証し，考察を行った．その結果，wildcard が有効であるデータベースとそうでないデータベースが存在するが，wildcard を許容した場合に精度の向上が見られることが多いという実験結果となった．これは，wildcard を許容することによる表現力の向上の結果であると言える．実際に，応用する際には，最小支持度の設定，冗長パターンの削除や頻出飽和・極大パターン集合の利用，wildcard 許容と許容なしの組み合わせなど，適切な設定を行うことで，精度の向上に務める必要がある．適切な設定や wildcard が有効なデータベースの性質などに対する研究が今後の課題である．

8. 謝辞

本研究は JSPS 科研費 26330242，26120503 および JST さきがけの助成を受けたものです.

参考文献

[Breiman 01] Breiman, L.: Random Forests, *Machine Learning*, Vol. 45, No. 1, pp. 5–32 (2001)

[Ho 98] Ho, T. K.: The Random Subspace Method for Constructing Decision Forests, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 20, No. 8, pp. 832–844 (1998)

[Karklin 05] Karklin, Y., Meraz, R. F., and Holbrook, S. R.: Classification of Non-Coding RNA Using Graph Representations of Secondary Structure, in *Biocomputing 2005, Proceedings of the Pacific Symposium, Hawaii, USA, 4-8 January 2005*, pp. 5–16 (2005)

[Kashima 03] Kashima, H., Tsuda, K., and Inokuchi, A.: Marginalized Kernels Between Labeled Graphs, in *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pp. 321–328 (2003)

[Kudo 05] Kudo, T., Maeda, E., and Matsumoto, Y.: An Application of Boosting to Graph Classification, in Saul, L. K., Weiss, Y., and Bottou, L. eds., *Advances in Neural Information Processing Systems 17*, pp. 729–736, MIT Press (2005)

[Rogers 10] Rogers, D. and Hahn, M.: Extended-Connectivity Fingerprints, *Journal of Chemical Information and Modeling*, Vol. 50, No. 5, pp. 742–754 (2010)

[Saigo 09] Saigo, H., Nowozin, S., Kadowaki, T., Kudo, T., and Tsuda, K.: gBoost: a mathematical programming approach to graph classification and regression, *Machine Learning*, Vol. 75, No. 1, pp. 69–89 (2009)

[Takigawa 13] Takigawa, I. and Mamitsuka, H.: Graph mining: procedure, application to drug discovery and recent advances, *Drug Discovery Today*, Vol. 18, No. 12, pp. 50 – 57 (2013)

[Wale 08] Wale, N., Watson, I. A., and Karypis, G.: Comparison of descriptor spaces for chemical compound retrieval and classification, *Knowl. Inf. Syst.*, Vol. 14, No. 3, pp. 347–375 (2008)

[岡崎 15] 岡崎 文哉, 瀧川 一学：Wildcard を許容した頻出部分グラフマイニング (情報論的学習理論と機械学習 情報論的学習理論ワークショップ (IBIS2015)), 電子情報通信学会技術研究報告 = IEICE technical report : 信学技報, Vol. 115, No. 323, pp. 25–32 (2015)

[瀧川 16] 瀧川 一学：多数のグラフからの統計的機械学習. 深化する機械学習技術の進展とその応用特集号, システム/制御/情報, Vol. 60, No. 3, pp. 107–112 (2016)

表 3: NCI データセットに対する実験結果

		frequent subgraphs		wildcard 1 許容		wildcard 2 許容		組み合わせ	
Database	minsup	出力数	ACC	出力数	ACC	出力数	ACC	出力数	ACC
NCI1	45%	73	0.62	569	0.674	2502	0.654		
	40%	102	0.584	814	0.663	3749	0.661		
	35%	139	0.617	1227	0.671	6090	0.659		
	30%	225	0.653	2076	0.656	10721	**0.686**		
	25%	323	0.623	3379	0.662	18965	0.665	10818	0.685
	20%	584	0.653	6352	0.675	37650	0.676	11079	0.659
	15%	1165	0.654	13944	0.659	NA	NA	11661	0.669
	10%	3283	0.673	43498	0.652	NA	NA	13778	**0.697**
	5%	27740	0.66	NA	NA	NA	NA	38235	0.674
NCI41	45%	81	0.564	645	0.6	2811	0.612		
	40%	109	0.616	893	0.599	4198	0.595		
	35%	151	0.602	1382	0.638	6917	0.613		
	30%	239	0.613	2261	0.633	11842	0.627		
	25%	371	0.63	3763	0.634	20948	0.64		
	20%	619	0.618	6853	0.651	40866	0.616		
	15%	1338	0.609	15642	0.652	NA	NA		
	10%	3576	0.606	46915	0.644	NA	NA		
	5%	24255	**0.657**	NA	NA	NA	NA		
NCI47	45%	74	0.597	576	0.666	2520	0.647		
	40%	104	0.611	827	0.644	3765	0.663		
	35%	139	0.65	1224	0.661	6100	**0.693**		
	30%	227	0.665	2080	0.656	10738	0.67	6188	0.676
	25%	327	0.663	3413	0.664	19144	0.69	6288	0.681
	20%	588	0.671	6432	0.683	38188	0.655	6549	0.653
	15%	1174	0.684	14085	0.657	NA	NA	7135	0.687
	10%	3290	0.665	43666	0.668	NA	NA	9251	0.692
	5%	26758	0.676	NA	NA	NA	NA	32719	0.685
NCI81	45%	67	0.592	509	0.612	2218	0.63		
	40%	93	0.61	741	0.621	3317	0.645		
	35%	126	0.626	1087	0.635	5298	0.644		
	30%	202	0.613	1825	0.628	9294	0.633		
	25%	296	0.598	3046	0.645	16800	**0.658**		
	20%	537	0.611	5741	0.645	33471	0.629	17040	0.639
	15%	1045	0.597	12285	0.643	NA	NA	17548	0.657
	10%	2857	0.64	37562	0.637	NA	NA	19361	0.656
	5%	28970	0.608	NA	NA	NA	NA	45474	0.617
NCI83	45%	75	0.552	590	0.588	2569	0.617		
	40%	104	0.56	827	0.601	3802	0.605		
	35%	135	0.583	1223	0.591	6080	0.605		
	30%	218	0.573	2026	0.618	10532	0.626		
	25%	340	0.562	3406	0.619	18863	0.601		
	20%	571	0.605	6243	0.594	36733	**0.639**		
	15%	1205	0.602	14042	0.608	NA	NA	37368	0.634
	10%	3294	0.597	42854	0.609	NA	NA	39457	0.625
	5%	22909	0.595	NA	NA	NA	NA	59072	0.631
NCI109	45%	71	0.608	536	0.665	2327	0.651		
	40%	99	0.637	775	0.618	3463	0.65		
	35%	131	0.626	1140	0.662	5546	0.656		
	30%	210	0.626	1891	0.67	9666	0.65		
	25%	301	0.617	3117	0.663	17437	0.684		
	20%	564	0.633	6014	0.67	35247	0.67		
	15%	1082	0.623	12984	0.68	NA	NA		
	10%	3104	0.677	41049	**0.695**	NA	NA		
	5%	28528	0.684	NA	NA	NA	NA	66473	0.68
NCI123	45%	60	0.618	446	0.667	1904	0.635		
	40%	83	0.608	652	0.631	2882	0.647		
	35%	118	0.635	969	0.648	4571	0.658		
	30%	183	0.638	1600	0.646	7991	0.647		
	25%	273	0.61	2740	0.643	14974	**0.674**		
	20%	499	0.622	5293	0.651	30655	0.652	15200	0.638
	15%	968	0.646	11468	0.634	72872	0.663	15669	0.662
	10%	2684	0.638	35407	0.632	NA	NA	17385	0.659
	5%	43926	0.617	NA	NA	NA	NA	58627	0.647
NCI145	45%	75	0.559	576	0.664	2521	0.653		
	40%	103	0.587	830	0.654	3793	0.645		
	35%	143	0.614	1260	0.664	6223	0.663		
	30%	231	0.622	2117	0.663	10896	0.67		
	25%	329	0.64	3449	0.669	19411	0.663		
	20%	600	0.647	6551	0.665	38895	0.67		
	15%	1188	0.645	14302	0.673	NA	NA		
	10%	3399	0.636	45064	**0.689**	NA	NA		
	5%	28166	0.683	NA	NA	NA	NA	69831	**0.708**
NCI167	45%	37	0.559	245	0.6	977	0.596		
	40%	48	0.588	330	0.603	1356	**0.631**		
	35%	61	0.574	461	0.596	2007	0.598	1369	0.591
	30%	91	0.587	703	0.595	3154	0.609	1399	0.607
	25%	116	0.587	1042	0.607	5340	0.606	1424	0.602
	20%	192	0.604	1877	0.61	10098	0.626	1500	0.6
	15%	357	0.59	3604	0.588	19841	0.602	1665	0.596
	10%	786	0.591	8780	0.615	53222	0.617	2094	0.597
	5%	3672	0.594	59157	0.615	NA	NA	4980	0.613
NCI220	45%	45	0.588	287	0.574	1071	0.575		
	40%	52	0.601	359	0.593	1495	0.589		
	35%	65	0.6	520	0.585	2290	0.586		
	30%	113	0.588	868	0.576	3865	0.602		
	25%	185	0.592	1536	0.601	7312	0.6		
	20%	322	0.599	3011	0.598	15466	0.599		
	15%	646	0.586	6725	0.59	37623	**0.605**		
	10%	1776	0.578	20844	0.591	NA	NA	40376	**0.618**
	5%	14961	0.579	NA	NA	NA	NA	65143	**0.648**
NCI330	45%	41	0.626	289	0.633	1194	0.646		
	40%	53	0.634	392	0.629	1740	0.62		
	35%	73	0.624	607	0.614	2822	0.661		
	30%	117	0.61	1001	0.662	4965	0.651		
	25%	199	0.599	1848	0.635	9620	0.66		
	20%	369	0.647	3602	0.662	19766	0.662		
	15%	775	0.633	8421	0.656	50272	0.664		
	10%	2367	0.678	30179	**0.688**	NA	NA		
	5%	NA	NA	NA	NA	NA	NA		