The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

進化計算を用いたフィルター特徴選択に関する評価指標の提案

Proposal a new feature evaluation measure for filter type feature subset selection with Evolutionary Algorithm

川村 篤志[*1]
Atsushi Kawamura

チャクラボルティ バサビ[*2]
Basabi Chakuraborty

[*1] 岩手県立大学大学院
Graduate School of Iwate Prefectural University

[*2] 岩手県立大学
Iwate Prefectural University

Feature Selection is an important preprocessing step for pattern recognition and data mining problems. This process selects necessary features and removes redundant features. In this work, we propose a new fitness function for feature subset evaluation. The proposed new fitness function is used for feature subset selection with EC algorithms. Simulation experiments using the benchmark datasets have been done and the results are compared with popular methods.

1. はじめに

近年，インターネットが普及し，SNS やクラウドからは画像やテキスト，ウェアラブル端末などの IoT デバイスからセンサーデータというように，様々なメディアからデータの送出が容易になってきている．

加えて，計算機性能や解析技術，ストレージ技術の向上により膨大なデータの収集や解析が容易になってきている．また，この膨大なデータを解析することで，知見を獲得しようという動きが活発化している．このため，データマイニングや機械学習の研究領域に対して重要性が高まってきている．

データの解析を行う際に，精度や処理速度を向上させ，効率的にデータを処理することが必要である．しかし，データをすべて用いて分類を行った場合，処理コストが高く，精度も悪くなってしまう傾向がある．これは，余計な特徴がデータ内に含まれているからである．そこで，解析を行う前処理として特徴選択という手法があげられる．これは非常に重要な処理であり，これを行うことで，精度の向上や処理コストの軽減が見込める．しかし，多くの特徴を持つデータから最適な特徴の組み合わせを選択することは，非常に困難である．したがって，効率よく最適であると思われる特徴部分集合を獲得できる手法を考える必要がある．

本稿では，特徴選択の評価指標に関して新たな評価指標の提案を行う．また，提案する評価指標と進化計算を用いて特徴選択を行い，全ての特徴を用いた場合と従来手法を用いた場合の精度と比較することで，本提案手法が有用であるかを検証することを目的とする．

2. 特徴選択

特徴選択とは，データから必要な特徴部分集合を選択し，不必要な特徴部分集合を削除する手法である．これを行うことによって，精度や汎化性能の向上，次元の呪いの緩和，処理の高速化などが期待される．

特徴選択は，主に 2 つの過程に分けられる．1 つ目は特徴部分集合を選択する探索手法である．どの特徴を使用し，どの特徴を削除するかを決定する方法である．例としては Sequential Forward Selection(SFS)，総当たり法，焼きなまし法，進化計算などがあげられる．特徴数を k とすると特徴部分集合の組み合わせ数は$2^k - 1$で指数的に増加するため，特徴数が多い場合，適切な特徴部分集合の選択は困難になる．

2 つ目は，1 つ目で得られた特徴部分集合をどのように評価するかという評価方法である．評価方法は，ラッパー法とフィルター法の 2 つに大別される．

ラッパー法は，実際に使用する分類器を特徴部分集合の評価に用いる手法である．そのため，高い精度を安定して得られる．また，どのデータにも宛がうことが可能であり，汎化性も高い．しかし，逐次的に選択される特徴部分集合すべてに対して分類器を用いて評価を行うため，計算コストが高く，過剰適合を起こす可能性がある．

フィルター法は，分類器の評価を使用せずデータの性質などによって評価を行う手法である．主に情報理論や統計的手法などの指標が用いられる．基本的に単純な評価式を用いているため，選択された特徴部分集合に対しての処理が高速である．しかし，データの種類や性質によって評価が左右されるため，あまり汎化的ではない．また，情報理論などのフィルター法は主に SFS，Sequential Backward Selection (SBS)などの探索手法と併用される．そのため，特徴数が多ければ多いほど評価する特徴部分集合の組み合わせはk^2で増大し，計算コストも大幅に増大していく．また，選択する n 個の特徴(以下 n)というパラメータを設定しなければならないためヒューリスティクスである．さらに，n が適切な設定であるかはわからない．

3. 進化計算

進化計算は，主に組み合わせ最適化問題の解決に用いられるメタヒューリスティクスな手法である．例として Genetic Algorithm(GA)[Goldberg 89] や Particle Swarm Optimization(PSO)などがあげられる．進化計算は，個体からなる個体群と反復される世代で指定された評価を最適化するものである．一般的には実数に対しての最適化に多く用いられるが，0 から 1 の二値で個体内部を表現したバイナリ型の進化計算も存在する．本研究では，GA と PSO をバイナリ型にした Binary Genetic Algorithm(BGA) と Binary Particle Swarm Optimization(BPSO)[Kennedy 97]を特徴部分集合の探索手法として用いる．

4. 提案手法

既存のフィルター法は，前述の通りnを設定しなければならないアルゴリズムがほとんどである．そして，特徴数が多ければ計

連絡先:川村 篤志，岩手県立大学大学院，岩手県滝沢市巣子 152-52，g231p007@s.iwate-pu.ac.jp

算コストも多くかかる. 本研究は, これらについて解決する新しい評価指標の提案である.

4.1 提案手法について

特徴選択には Consistency という一貫性指標がフィルター法で存在している. 本研究では, 自然言語処理における各クラスタの一貫性を示す Coherence を数値データに使用できないかと考えた. そのため, Tf-idf で求められる文書中に存在する各単語の重要度の行列, 各クラスタにおける重要単語間の数値を求める UCI Coherence[Newman 10]を用いて評価を行う. それに加え各クラスタ間の距離関数として Earth Mover's Distance(EMD) [Rubner 00]を用いる.

4.2 提案手法の過程

データを 0 から 1 で正規化を行う. 0 から 1 をビンの数で分割しデータをスケーリングする. スケーリングされたデータを重要度行列に変換する. この時の重要度行列は, サンプル数 - 要素数の行列である. この重要度行列に対して Non-negative Matrix Factorization(NMF)を行う. 重要度行列を X とすると, W×H の行列に分解される. W はサンプル数-特徴行列, H は特徴-要素列である. NMF で抽出される特徴をクラス数とした場合. 各クラスタについて要素ごとの重要度が得られるので, 各クラスタの重要度上位 n 個の要素を取り出しそれぞれ Pointwise Mutual Information(PMI)を求めることで Coherence を算出する.

4.3 提案手法の説明

データのサンプル数を NS, クラス数を NC, 要素を x とする. Coherence は式(1)によって求められ, EMD は式(7)によって求められる. そして, 各クラスの Coherence の平均をS_1, 各クラスタ間の EMD の平均をS_2とした場合, S は式(11)で算出される. 本研究では, この S を進化計算によって最大化する. また, 進化計算により選択される n はメタヒューリスティクスに選択されるため n を設定しなくてよい.

$$\mathrm{Coherence} = \sum_{i,j} \mathrm{Score}(x_i, \mathrm{x_j}) \quad \ldots(1)$$

$$\mathrm{Score}_{\mathrm{UCI}(\mathrm{x_i},\mathrm{x_j})} = NPMI(x_i, x_j) \quad \ldots(2)$$

$$\mathrm{NPMI}_{(\mathrm{x_i},x_j)} = \frac{PMI(x_i, x_j)}{-\log_2(P(x_i, x_j))} \quad \ldots(3)$$

$$\mathrm{PMI}(\mathrm{x_i}, \mathrm{x_j}) = \log_2\left(\frac{P(x_i, x_j)}{P(x_i)P(x_j)}\right) \quad \ldots(4)$$

$$\mathrm{P}(\mathrm{x_i}) = \frac{NS(x_i)}{|NS|} \quad \ldots(5)$$

$$\mathrm{P}(\mathrm{x_i}, \mathrm{x_j}) = \frac{NS(x_i, x_j)}{|NS|} \quad \ldots(6)$$

$$\mathrm{EMD}(\mathrm{P}, \mathrm{Q}) = \min \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij} \quad \ldots(7)$$

NCombi は各クラスタ間の組み合わせである.

$$\mathrm{NCombi} = NC \times \frac{(NC-1)}{2} \quad \ldots(8)$$

$$\mathrm{S_1} = \frac{1}{NC} \sum_{c=1}^{NC} Coherence\,(\mathrm{c_i}) \quad \ldots(9)$$

$$\mathrm{S_2} = \frac{1}{NCombi} \sum_{i<j}^{NCombi} EMD(C_i, C_j) \quad \ldots(10)$$

$$\mathrm{S} = \mathrm{S_1} \times \sqrt{S_2} \quad \ldots(11)$$

5. 実験

本研究で提案した指標について, 様々なデータセットを用いて従来手法等の比較評価を行う. また, 得られた結果に対して考察を行う.

5.1 評価

実験での評価を行うベンチマークデータセットとして UCI Datasets Repository[Dua 17] , NIPS2003[NIPS 03] , WCCI2006[WCCI 06]から合わせて 12 個のデータセットを使用する. 表 1 には, データ名・特徴数・クラス数・データ数を表記する.

比較評価として, フィルター法から既存のアルゴリズムをいくつか用いる. 各データセットに対して, 特徴数を 10 で除算し, 評価の高い順にソートした特徴の添字に対して上位 10 分の 1 から 10 分の 9 までそれぞれを評価する.

実験における最終評価は LinearSVM で 10 fold cross validation を行った分類精度の平均を用いる.

実験において, 全ての検証を同じ環境で行い, 使用したアルゴリズム全てのタイムアウト時間を 48 時間とした.

表 1. データセット詳細

データ名	特徴数	クラス数	データ数
WINE	13	3	178
CANCER	30	2	569
ADA	48	2	Tr : 4147 Te : 415
SONAR	60	2	208
GAS	128	6	Tr : 6955 Te : 6955
SYLVA	216	2	Tr : 13086 Te : 1308
MADELON	500	2	Tr : 2000 Te : 600
GINA	970	2	Tr : 3153 Te : 315
HIVA	1617	2	Tr : 3845 Te : 384
ARCENE	10000	2	Tr : 100 Te : 100
DEXTER	20000	2	Tr : 300 Te : 300
DOROTHEA	100000	2	Tr : 800 Te : 350

5.2 手法

提案手法の各パラメータについては, ビンを 100, 分解するクラスタ数はデータのクラス数とし, 各クラスタの重要度上位 5 個を抽出とした. 探索手法として用いた進化計算のパラメータは, 表 2 に表記する.

従来手法として, フィルター法でよく用いられている Minimum redundancy maximum relevance(mRMR) [Peng 05][Brown 12]と Correlation Feature Selection(CFS) [Mark 99], 2 つのアルゴリズムを比較として用いる.

表 2. 進化計算パラメータ

アルゴリズム	群数	反復数	パラメータ
BGA	20	300	二点交叉 ランク法 mutation=0.05
BPSO	20	300	w=0.5 c=1.0

5.3 結果

比較として, 全ての特徴を使用して評価した結果を表 3 に表記する. 各データセットにおけるフィルター法の評価値と特徴数の一部を表 4 に表記する. また, 各データセットに対してフィルター法での精度が最大だった部分を選択された特徴数とともに表 5, 6 に表記する. 斜線は前述にあるタイムアウトを表す.

提案手法の精度と選択された特徴数を表 7, 8 に表記する.

表 3. 全特徴使用時の精度

データ名	精度
WINE	0.885185
CANCER	0.901075
ADA	0.739277
SONAR	0.746031
GAS	0.776851
SYLVA	0.980199
MADELON	0.538333
GINA	0.790791
HIVA	0.937240
ARCENE	0.830000
DEXTER	0.933333
DOROTHEA	0.931429

表 4. Cancer

	mRMR	CFS
使用添字割合	精度	
1/10	0.811891	0.649123
2/10	0.847953	0.901559
3/10	0.896686	0.919103
4/10	0.872320	0.927875
5/10	0.916179	0.922027
6/10	0.867446	0.896686
7/10	0.873294	0.928850
8/10	0.892788	0.905458
9/10	0.882066	0.877193

表 5. mRMR の結果

	精度	使用添字割合	特徴数
WINE	0.854938	9/10	11
CANCER	0.916179	5/10	15
ADA	0.768675	2/10	9
SONAR	0.809524	7/10	42
GAS	0.802181	9/10	115
SYLVA	0.984072	3/10	64
MADELON	0.53	8/10	400
GINA	0.799471	7/10	679
HIVA	0.966146	1/10	161
ARCENE	0.84	7/10	7000
DEXTER			
DOROTHEA			

表 6. CFS の結果

	精度	使用添字割合	特徴数
WINE	0.91358	9/10	11
CANCER	0.92885	7/10	21
ADA	0.781928	2/10	9
SONAR	0.793651	6/10	36
GAS	0.979415	5/10	64
SYLVA	0.981269	5/10	108
MADELON	0.536111	7/10	350
GINA			
HIVA			
ARCENE			
DEXTER			
DOROTHEA			

表 7. 提案指標＋BGA

データ名	精度	特徴数
WINE	0.896667	10.80
CANCER	0.917778	24.60
ADA	0.730040	29.33
SONAR	0.783069	52.33
GAS	0.764411	105.8
SYLVA	0.932263	149.0
MADELON	0.518000	237.5
GINA	0.801746	898.0
HIVA	0.950521	478.0

ARCENE	0.810000	451.0
DEXTER	0.803333	6116.5
DOROTHEA	0.891429	1358.5

表 8.　提案指標＋BPSO

データ名	精度	特徴数
WINE	0.834656	10.43
CANCER	0.903509	19.00
ADA	0.719663	24.40
SONAR	0.730159	40.20
GAS	0.766783	75.33
SYLVA	0.959531	116.67
MADELON	0.517500	260.67
GINA	0.797249	897.67
HIVA	0.937847	855.67
ARCENE	0.822750	492.25
DEXTER	0.908889	9874.0
DOROTHEA	0.881905	2002.33

5.4 考察

実験の結果，全特徴を用いた場合と比較すると，精度の上昇が見られたデータセットもあったが，変化がほとんど見られないデータセットや減少をしているデータセットも確認された．しかし，ほとんどのデータセットは微小な減少に留まっている．さらに，選択された特徴数については，ほとんどのデータセットに対して削減が行われている．

前述の通り，フィルター法は精度がデータセットによって左右されることも多くすべてのデータセットに対して精度を向上させることは難しい．また，進化計算に関しても汎用的な探索を行うため，ヒューリスティクスよりも精度が劣る傾向がある．

大きなデータに対して特徴数の影響をあまり受けることなく計算が可能であることが確認された．

6. おわりに

6.1 まとめ

本研究では，特徴選択のフィルター法に関して，Coherence と EMD を用いて独自の評価指標を提案した．結果として，多くのデータセットに対し良好な精度を獲得できることが確認された．

n を設定しなくてもよいこと，大きなデータセットに対しても作動することからも，本研究は有用であるということがいえる．

6.2 今後の展望

今後は，より多くのデータに適用できる指標にするため，教師無しのデータに対しても特徴選択が行えるように改良を加える．

また，今回は自然言語処理の UCI Coherence を求める過程に則って算出を行う評価指標を提案したが，様々な過程が存在するため，安定した精度，多様なデータへの汎用性の向上などを今後の目標とし，さらなる調整を行う．

参考文献

[Goldberg 89] Goldberg, David.:Genetic Algorithms in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley Professional, (1989)

[Kennedy 97] Kennedy, J. and Eberhart, R. C.: A discrete binary version of the particle swarm algorithm, Conference on Systems, Man, and Cybernetics, Piscataway. NJ: IEEE Service Center, pp. 4104-4109 (1997)

[Newman 10] David, Newman.;Jey, Han.;Lau, Karl, Grieser.; and imothy Baldwin. "Automatic evaluation of topic coherence". HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics Pages 100-108. (2010)

[Rubner 00] Rubner, Y, Tomasi, C, and Guibas, I: "The Earth Mover's Distance as a Metric for Image Retrieval", International Journal of Computer Vision, 14(3), pp.130-137, (2000).

[Dua 17] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[WCCI 06] IEEE World Congress on Computational Intelligence 2006: Performance prediction challenge. , http://clopinet.com/isabelle/Projects/modelselect/WCCI06challengeResu.html

[NIPS 03] Neural Information Processing Systems Conference: Feature selection challenge , http://clopinet.com/isabelle/Projects/NIPS2003/ , (NIPS 2003)

[Peng 05] Peng, H. C.; Long, F.; Ding, C. "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy". IEEE Transactions on Pattern Analysis and Machine Intelligence. 27 (8): 1226–1238. (2005).

[Brown 12] Brown, Gavin et al. "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection.", In the Journal of Machine Learning Research(JMLR), (2012).

[Mark 99] M. Hall, "Correlation-based Feature Selection for Machine Learning", (1999).