MINI-BATCH NORMALIZED MUTUAL INFORMATION: A HYBRID FEATURE SELECTION METHOD

Thejas G. S.[1], S. R. Joshi [2], S. S. Iyengar[1], N. R. Sunitha[2] , Prajwal Badrinath[1]
[1]Florida International University, [2]Siddaganga Institute of Technology, India

Problem

- To find optimized feature subset
- Literature: Existing methods have their own performance limitations

Summary of Contribution

- Propose Hybrid Feature Selection methods
  - **Method 1:** MiniBatch K-Means Normalized Mutual Information Feature Inclusion (KNFI)
  - **Method 2:** MiniBatch K-Means Normalized Mutual Information least ranked Feature Exclusion (KNFE)

Our Approach

- 2 phase Hybrid Feature Selection methods
  - Combination of filter-wrapper approach
    - Feature ranking function based on the filter approach
      - Cluster the features based upon the total classes in the dataset
      - Cluster quality – Normal Mutual Information score between 0 to 1
      - Higher the rank score better the classification
    - Selection of optimal features based upon the rankings
      - Feature Inclusion (KNFI) – Highest ranked feature considered initially
      - Least Ranked Feature Exclusion (KNFE) – initially, all features and classification accuracy are considered and then elimination begins

Experiment

- 15 datasets – 9 binary class, 6 Multi class
- Base classifier: Random Forest
- Evaluation parameters
  - Accuracy, Precision, Recall, F1 Score, and Area Under Curve (AUC)
- Work compared
  - Recursive Feature Elimination (RFE)[1] and other works [2, 3, 4, 5, 6, 7, 8]

Results – Binary Class

Method	Ftr	Acc	AUC	F1
AF	43	99.93	99.46	99.93
KNFI	17	**99.963**	99.614	99.96
RFE	17	99.960	99.612	99.96
KNFE	-6	**99.944**	99.96	99.94

UNSW NB15 Dataset

Method	Ftr.	Acc	AUC	F1
AF	9	99.9179	99.9179	99.92
KNFI	4	**99.919**	99.919	99.92
RFE	4	99.919	99.919	99.92
KNFE	0	**99.9179**	99.917	99.92

Talking Dataset Version 2

Method	Ftr	Acc	AUC	F1
AF	34	92.96	90.91	92.84
KNFI	6	**97.18**	95.23	97.14
RFE	6	91.54	7.92	91.55
KNFE	-7	**95.77**	94.238	95.74

Ionosphere Dataset

Method	Ftr.	Acc	AUC	F1
AF	57	98.04	97.69	98.04
KNFI	15	**97.82**	97.52	97.82
RFE	15	97.285	96.69	97.27
KNFE	-3	**98.58**	98.301	98.93

Spambase Dataset

Method	Ftr.	Acc	AUC	F1
AF	25	83.029	54.235	77.89
KNFI	7	**83.4375**	53.283	77.89
RFE	7	83.075	53.013	77.63
KNFE	-17	**83.381**	52.456	77.36

Avazu Dataset

Method	Ftr.	Acc	AUC	F1
AF	60	92.86	93.05	92.88
KNFI	3	**95.24**	95.138	95.24
RFE	3	88.09	88.88	88.16
KNFE	-9	**97.62**	97.91	97.63

Sonar Dataset

Method	Ftr.	Acc	AUC	F1
AF	8	95.127	91.672	95.08
KNFI	6	**94.252**	90.434	94.14
RFE	6	94.784	91.059	94.67
KNFE	0	**95.20**	91.72	95.11

Talking dataset Version 1

Method	Ftr.	Acc	AUC	F1
AF	39	73.545	62.386	70.29
KNFI	3	**70.205**	57.725	65.85
RFE	3	70.268	55.902	63.85
KNFE	-5	**73.545**	62.45	70.33

Criteo Dataset

Method	Ftr.	Acc	AUC	F1
AF	10	98.540	98.113	98.53
KNFI	4	**97.810**	97.517	97.81
RFE	4	94.890	93.744	94.84
KNFE	-3	**98.540**	98.113	98.53

Breast Cancer Dataset

Results – Multi Class

Method	Ftr.	Acc	F1
AF	43	89.326	88.87
KNFI	16	**90.107**	88.88
RFE	16	89.356	89.02
KNFE	-18	**89.591**	89.02

UNSW NB15 Dataset

Method	Ftr.	Acc	F1
AF	4	96.666	96.67
KNFI	2	**99.9999**	99.99
RFE	2	99.999	99.999
KNFE	-4	**99.999**	99.999

Iris Dataset

Method	Ftr.	Acc	F1
AF	56	33.333	37.78
KNFI	3	**66.666**	68.25
RFE	3	50.00	52.78
KNFE	-14	**66.666**	68.25

Lung Cancer Dataset

Method	Ftr.	Acc	F1
AF	13	41.667	34.60
KNFI	4	**56.667**	51.53
RFE	4	43.333	36.71
KNFE	-11	**51.667**	40.90

Heart Disease Dataset

Method	Ftr.	Acc	F1
AF	18	86.66	85.19
KNFI	2	**90.00**	89.78
RFE	2	76.66	80.00
KNFE	-2	**86.66**	75.17

Lymphography Dataset

Method	Ftr.	Acc	F1
AF	8	24.521	22.86
KNFI	1	**21.650**	20.27
RFE	1	17.344	17.14
KNFE	0	**25.239**	23.61

Abalone Dataset

Results – Other Works

Method	# Ftr.	F1	RC	Pr	Acc
Venkatesh *et al.*[3]	15	95.09	94.65	95.70	95.28
HGEFS [2]	n.a.	n.a.	n.a.	n.a	91.33
FSFOA [7]	n.a.	n.a.	n.a.	n.a	95.12
KNFI	6	97.14	97.18	97.29	**97.18**
KNFE	-7	95.74	95.77	95.76	**95.77**

Comparisons of Ionosphere data with Previous Studies

Ftr. selection method	# features	accuracy
GAMIFS[6]	3	83.50
NMIFS[6]	3	75.8
MIFS[4]	3	78.4
MIFS-U[5]	3	81.2
OFS-MI [8]	3	78.4
KNFE	3	**84.15**
KNFI	15	**97.82**
KNFE(MAX)	54	**98.59**

Comparisons for Spambase dataset with previous studies

Ftr. selection method	# features	accuracy
NMIFS[6]	15	86.73
MIFS(β=0.5)[4]	15	85.96
MIFS-U(β = 0.5)[5]	15	84.04
HGEFS[2]	N.A.	83.00
FSFOA[7]	N.A	86.98
KNFE	15	**92.85**
KNFI	3	**95.24**
KNFE(MAX)	51	**97.62**

Comparisons for Sonar dataset with previous studies

Conclusions

- Proposed hybrid method utilizes the advantages of both filter and wrapper
- No constraint for the user to input the number of features required as in RFE
- NMI as a metric to rank the features after clustering by Mini-Batch K-Means
  - Mini-Batch k-mean is faster than K-mean

https://github.com/thejasgs

Acknowledgement

This work was supported by the Florida International University Graduate School Dissertation Year Fellowship Award received by the author G. S. Thejas.

References

1. Fabian Pedregosa, Ga¨el Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. Journal of machine learning research , 12(Oct):2825–2830, 2011.
2. X. Xue, M. Yao, and Z. Wu, “A novel ensemble-based wrapper method for feature selection using extreme learning machine and genetic algorithm,” Knowledge and Information Systems, vol. 57, no. 2, pp. 389–412, 2018.
3. B. Venkatesh and J. Anuradha, “A hybrid feature selection approach for handling a high-dimensional data,” in Innovations in Computer Science and Engineering. Springer, 2019, pp. 365–373.
4. R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” IEEE Transactions on neural networks, vol. 5, no. 4, pp. 537–550, 1994.
5. N. Kwak and C.-H. Choi, “Input feature selection for classification problems,” IEEE transactions on neural networks, vol. 13, no. 1, pp. 143–159, 2002.
6. P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, “Normalized mutual information feature selection,” IEEE Transactions on Neural Net-works, vol. 20, no. 2, pp. 189–201, 2009.
7. Manizheh Ghaemi and Mohammad-Reza Feizi-Derakhshi. Feature selection using forest optimization algorithm. Pattern Recognition , 60:121–129, 2016.
8. T. W.S. Chow and D. Huang. Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. Trans. Neur. Netw. , 16(1):213–224, January 2005.