The Role of Biomedical Dataset in Classification

Ajay Kumar Tanwani and Muddassar Farooq

Next Generation Intelligent Networks Research Center (nexGIN RC)
National University of Computer & Emerging Sciences (FAST-NUCES)
Islamabad, 44000, Pakistan
{ajay.tanwani,muddassar.farooq}@nexginrc.org

Abstract. In this paper, we investigate the role of a biomedical dataset on the classification accuracy of an algorithm. We quantify the complexity of a biomedical dataset using five complexity measures: correlation-based feature selection subset merit, noise, imbalance ratio, missing values and information gain. The effect of these complexity measures on classification accuracy is evaluated using five diverse machine learning algorithms: J48 (decision tree), SMO (support vector machines), Naive Bayes (probabilistic), IBk (instance based learner) and JRIP (rule-based induction). The results of our experiments show that noise and correlation-based feature selection subset merit – not a particular choice of algorithm – play a major role in determining the classification accuracy. In the end, we provide researchers with a meta-model and an empirical equation to estimate the classification potential of a dataset on the basis of its complexity. This well help researchers to efficiently pre-process the dataset for automatic knowledge extraction.

Keywords: Classification, Complexity Measures, Biomedical Datasets.

1 Introduction

A diverse set of machine learning and data mining algorithms have been proposed to extract useful information from a dataset. But, the learning behavior of all these algorithms is dependent on the complexity of underlying data [1]. Biomedical datasets, in this context, pose a unique challenge to machine learning techniques for classification because of their high dimensionality, multiple classes, noisy data and missing values [2]. Therefore, we advocate the need to separately study the impact of the complexity of biomedical dataset in classification.

In this paper, we systematically investigate the role of biomedical dataset in classification using a number of complexity measures and a diverse set of machine learning algorithms. The empirical study is performed on 31 biomedical datasets publicly available from the UCI Machine Learning repository [3]. The goal is to resolve the uncertainties associated with the complexity of biomedical dataset and the resulting accuracy of classification. The outcome of our study is a novel framework to estimate the classification potential of a biomedical dataset. This will prove useful in understanding the nature of a biomedical dataset for efficient pre-processing and automatic knowledge extraction.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 370–374, 2009.

2 Complexity Measures of Biomedical Datasets

1) Correlation-Based Feature Selection Subset Merit - CfsSubset Merit. Correlation-based feature selection (Cfs) is used to select a subset of attributes that are highly correlated with the class but have low inter-correlation. The correlation between the two attributes A and B with entropies $H(A)$ and $H(B)$ is measured using the *symmetrical uncertainty* [4]:

$$U(A,B) = 2\frac{H(A) + H(B) - H(A,B)}{H(A) + H(B)} \tag{1}$$

where H(A,B) is the joint entropy calculated from the joint probabilities of all combination of attribute values. The merit of a subset formed with correlation-based feature selection M_{cfs} is calculated using [4]:

$$M_{cfs} = \frac{\sum_{i=1}^{N_a} U(A_j, C)}{\sqrt{\sum_{i=1}^{N_a} \sum_{j=1}^{N_a} U(A_i, A_j)}} \tag{2}$$

where N_a is the number of attributes in the set and C is the class attribute. The CfsSubset merit provides a measure of the quality of attributes in a dataset.

2) Noise. Brodley and Friedl characterized noise as the proportion of incorrectly classified instances by a set of trained classifiers [5]. We quantify noise as the sum of all off-diagonal entities where each entity is the minimum of all the corresponding elements in a set of confusion matrices. The advantage of our approach is that we separately identify misclassified instances of every class and only categorize those as noisy which are misclassified by all the classifiers. The percentage of class noise N_o in a dataset with I_n instances can be computed as below:

$$N_o = \left(\frac{1}{I_n} \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} min(C_1(i,j), C_2(i,j)......C_n(i,j)) \right) 100 \qquad (i \neq j) \tag{3}$$

where C_n is the n^{th} confusion matrix in a set of n classifiers, N_c is the number of classes, $min(C_1(i,j), C_2(i,j)......C_n(i,j))$ is an entity for corresponding i and j that represents minimum number of class instances misclassified by all the classifiers. We have used the same set of classifiers as used for our comparative study to determine percentage of noise levels in the datasets. It is evident from Table 1 that biomedical datasets are generally associated with high noise levels.

3) Imbalance Ratio. We propose the following definition of imbalance ratio I_r to cater for proportion of all class distributions in a dataset:

$$I_r = \frac{N_c - 1}{N_c} \sum_{i=1}^{N_c} \frac{I_i}{I_n - I_i} \tag{4}$$

where I_r is in the range ($1 \leq I_r < \infty$) and I_r = 1 is a completely balanced dataset having equal instances of all classes. I_i is the number of instances of i^{th} class and I_n is the total number of instances.

Table 1. The Table shows: (1) Complexity of datasets quantified in terms of CfsSubset Merit (CfsSub Merit), Noise, Imbalance Ratio (Imb Ratio), Average Information Gain (Info Gain) and Missing Values; (2) Classification accuracies of compared algorithms; bold entry in every row represents the best accuracy

Dataset	**Complexity of Dataset**					**Classifiers**					**Mean**
	CfsSub Merit	**Noise**	**Imb Ratio**	**Info Gain**	**Missing Values**	**J48**	**SMO**	**NB**	**IB*k***	**JRIP**	
Ann-Thyroid	0.64	0.11	8.37	0.037	0	**99.69**	93.79	95.42	94.12	99.53	96.51 ± 2.89
Breast Cancer	0.70	2.72	1.21	0.451	0.23	94.56	96.71	95.99	**96.99**	95.42	95.94 ± 0.98
Breast Cancer Diagnostic	0.67	2.11	1.14	0.303	0	92.97	**97.89**	92.62	97.19	93.67	94.87 ± 2.48
Breast Cancer Prognostic	0.12	13.64	1.76	0.004	0.06	73.74	**75.76**	67.17	73.23	**75.76**	73.13 ± 3.52
Cardiac Arrhythmia	0.47	11.28	1.57	0.047	0.32	65.49	**70.57**	61.73	58.85	69.25	65.18 ± 4.94
Cleveland-Heart	0.26	17.82	1.37	0.115	0.15	55.45	**59.08**	55.45	**59.08**	53.14	56.43 ± 2.59
Contraceptive Method	0.08	31.98	1.05	0.041	0	52.14	48.20	50.78	48.47	**52.41**	50.40 ± 1.98
Dermatology	0.77	0.82	1.05	0.442	0.06	93.99	95.35	**97.27**	95.63	86.88	93.82 ± 4.05
Echocardiogram	0.65	6.06	1.24	0.084	4.67	**90.84**	86.26	87.02	86.26	**90.84**	88.24 ± 2.39
E-Coli	0.67	6.55	1.25	0.678	0	84.23	**87.20**	85.12	86.01	81.25	84.76 ± 2.25
Haberman's Survival	0.08	16.67	1.57	0.023	0	71.89	73.53	**74.84**	69.28	72.22	72.35 ± 2.07
Hepatitis	0.32	10.97	2.0	0.058	5.67	81.94	**87.10**	83.22	85.16	76.77	82.84 ± 3.91
Horse Colic	0.32	11.96	1.14	0.061	19.39	85.33	83.42	80.43	81.79	**86.96**	83.59 ± 2.62
Hungarian Heart	0.27	13.61	1.74	0.079	20.46	**68.71**	68.02	65.31	66.33	64.63	66.60 ± 1.74
Hyper Thyroid	0.30	0.34	28.81	0.012	2.17	**98.94**	97.77	95.39	97.83	98.49	97.68 ± 1.37
Hypo-Thyroid	0.42	0.54	9.99	0.024	6.74	**99.24**	97.44	97.91	97.28	**99.24**	98.22 ± 0.96
Liver Disorders	0.06	9.86	1.05	0.011	0	**68.70**	58.26	55.36	59.13	64.64	61.22 ± 5.36
Lung Cancer	0.50	21.88	1.02	0.152	0.28	50.00	40.62	**62.50**	40.62	43.75	47.50 ± 9.22
Lymph Nodes	0.41	10.81	1.46	0.138	0	77.03	**86.49**	83.11	83.78	76.35	81.35 ± 4.45
Mammographic Masses	0.33	14.15	1.01	0.193	3.37	82.73	78.88	83.14	80.12	**83.25**	81.62 ± 1.99
New Thyroid	0.74	2.79	1.78	0.602	0	92.09	89.77	**96.74**	93.95	93.02	93.12 ± 2.56
Pima Indians Diabetes	0.16	20.18	1.20	0.064	0	73.83	**77.34**	76.30	73.18	75.13	75.16 ± 1.72
Post Operative Patient	0.05	30.00	1.90	0.016	0.44	**70.00**	**70.00**	67.78	68.89	**70.00**	69.33 ± 0.99
Promoters Genes	0.36	4.72	1.00	0.078	0	81.13	**93.40**	90.57	79.24	83.02	85.47 ± 6.17
Protein Data	0.02	45.48	1.19	0.065	0	**54.52**	**54.52**	**54.52**	**54.52**	**54.52**	54.52 ± 0.00
Sick	0.42	0.71	7.72	0.013	2.24	**98.75**	93.86	92.68	95.96	98.21	95.89 ± 2.65
Splice-Junction Genes	0.48	4.60	1.15	0.022	0	93.05	91.70	**93.90**	79.80	93.20	90.33 ± 5.94
Statlog Heart	0.32	15.19	1.02	0.092	0	76.67	82.59	**84.81**	81.11	77.04	80.44 ± 3.54
Switzerland Heart	0.09	32.52	1.14	0.023	17.07	29.27	39.02	35.77	30.89	**39.84**	34.96 ± 4.74
Thyroid0387	0.42	1.35	2.99	0.091	5.5	**95.76**	77.77	78.42	81.81	93.93	85.54 ± 8.65
VA-Heart	0.07	27.00	1.04	0.023	26.85	34.00	**35.00**	34.00	32.00	30.00	33.00 ±2.00
Mean	**0.36**	**12.53**	**2.97**	**0.13**	**3.73**	76.99$^{(2)}$ ± 19.02	**77.01**$^{(1)}$ ± **18.62**	76.62$^{(3)}$ ± 18.28	75.11$^{(5)}$ ± 19.34	76.53$^{(4)}$ ± 18.82	

4) Missing Values. The datasets obtained from clinical databases contain several missing fields as their inherent characteristic. Therefore, we quantify the percentage of missing values in a dataset to study their effect on classification accuracy.

5) Information Gain. Information gain is an information-theoretic measure that evaluates the quality of an attribute in a dataset based on its entropy [4]. We use the average information gain of a dataset to give a measure of the quality of its attributes for classification.

3 Experiments, Results and Discussions

We now present the results of our experiments that we have done to analyze the complexity of 31 biomedical datasets with five different algorithms: J48, SMO, Naive Bayes, IBk and JRIP [2]. We have used the standard implementations of these schemes in Wakaito Environment for Knowledge Acquisition (WEKA) [4] to remove any customized bias in our study. A careful insight into the results in Table 1 helps to draw an important conclusion: ***the variance in accuracy of classifiers on a particular dataset is significantly smaller compared with the variance in accuracy of the same classifier on different datasets***. This implies that the nature of dataset has a very strong impact on the classification accuracy of a dataset compared to the choice of classifier. In Figure 1, we present the effect of our complexity measures with mean classification accuracy of all algorithms. It is obvious that the percentage of noise in a dataset effectively determines the classification accuracy; the CfsSubset merit is directly proportional to the classification accuracy; the high average information gain of a dataset yields better classification accuracy; the high percentage of missing values significantly degrade the classification accuracy while the imbalance ratio in a dataset has a minor impact on the resulting accuracy.

In order to get better insights in Figure 1, a meta-dataset is created consisting of our five complexity measures as its attributes. The output classification potential of a dataset is categorized into three classes depending upon the classification accuracy: good (greater than 85%), satisfactory (65-85%) and bad (less than 65%). The classification models are extracted using JRIP and J48 with resulting classification accuracies of 80.64% and 77.42%. The meta-model shows that noise and CfsSubset Merit are the two most important attributes in estimating the classification potential of a dataset.

Classification Rules of JRIP

```
(noise >= 17.82) => class=bad (8.0/2.0)
(noise <= 6.06) => class=good (12.0/0.0)
 => class=satisfactory (11.0/1.0)
```

Decision Tree of J48

```
noise <= 4.72: good (12.0)
noise > 4.72
|   CfsSubset <= 0.263: bad (9.0/1.0)
|   CfsSubset > 0.263: satisfactory (10.0/1.0)}
```

To generalize the findings of our study, we map an equation on the obtained results of noise (N_o) and CfsSubset Merit (M_{cfs}) to determine the classification potential (C_P)

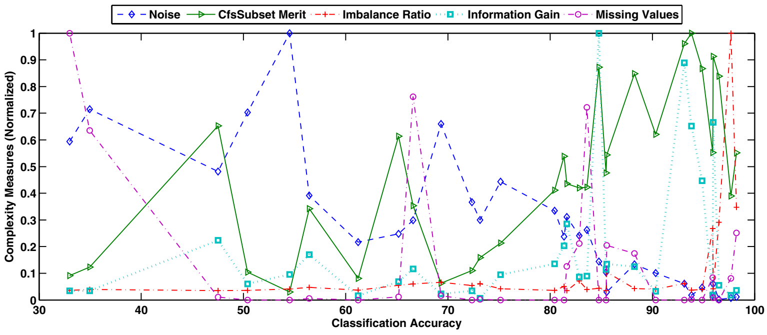

Fig. 1. Effect of Complexity Measures on Classification Accuracy

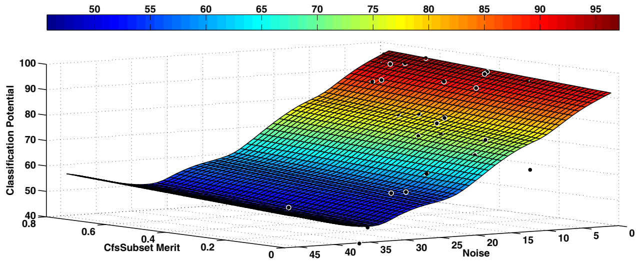

Fig. 2. Classification Potential as Function of Noise and CfsSubset Merit

of a dataset (see Figure 2). The equation is obtained using fitness criteria of lowest sum of squared absolute error:

$$C_P = 98.66 - 1.22 * N_o - \frac{1.43}{M_{cfs}} - 0.06 * {N_o}^2 - \frac{0.09}{{M_{cfs}}^2} + 0.19 * \frac{N_o}{M_{cfs}} \qquad (5)$$

4 Conclusion

In this paper, we have quantified the complexity of a biomedical dataset in terms of correlation-based feature subset merit, noise, imbalance ratio, missing values and information gain. The effect of complexity on classification accuracy is evaluated using five well-known diverse algorithms. The results show that the complexity measures – noise and CfsSubset merit – predominantly determines the classification accuracy of a biomedical dataset rather than the choice of a particular classifier. The major contribution of this paper is a novel methodology for estimating the classification potential of a dataset using its complexity measures.

References

1. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 289–300 (2002)
2. Tanwani, A.K., Afridi, J., Shafiq, M.Z., Farooq, M.: Guidelines to select machine learning scheme for classifcation of biomedical datasets. In: EVOBIO 2009. LNCS, vol. 5483, pp. 128–139. Springer, Heidelberg (2009)
3. UCI repository of machine learning databases, University of California-Irvine, Department of Information and Computer Science, `http://www.ics.uci.edu/~mlearn/MLRepository.html`
4. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
5. Brodley, C.E., Friedl, M.A.: Identifying mislabeled training data. Journal of Artificial Intelligence Research 11, 131–167 (1999)