Low Dimensionality or Same Subsets as a Result of Feature Selection: An In-Depth Roadmap

Antonio J. Tallón-Ballesteros(✉) and José C. Riquelme

Department of Languages and Computer Systems,
University of Seville, Seville, Spain
atallon@us.es

Abstract. This paper addresses the situation that may happen after the application of feature subset selection in terms of a reduced number of selected features or even same solutions obtained by different algorithms. The data mining community has been working for a long time with the assumption that meaningful attributes are either highly correlated with the class or represent a consistent subset, that is, with no inconsistencies. We have analysed around a hundred data sets very varied with a number of attributes below one hundred, a number of instances not greater than fifty thousand and a number of classes below fifty. Basically, in the first round we applied two different feature subset selection methods to pick up the figures in terms of reduced dimensionality. After that, we divided them into different groups according to the number of selected attributes. Next, we deepened the analysis in every category and we added a new feature selection procedure. Finally, we assessed the performance of the original problem and the reduced subsets with four classifiers providing some prospective directions.

Keywords: Classification · Feature subset selection · Correlation · Consistency · Feature ranking

1 Introduction

Predictive data analytics [8] encompasses the business and data processes and computational models that enable a business to make data-driven decisions. The key idea is to move from data to insights to decisions. Machine learning algorithms work by searching through a set of possible prediction models for the model that best captures the relationship between the descriptive features and the target feature [22]. An obvious search criteria to drive this search is to look for models that are consistent with the data. The classifier training could achieve to three types of models: an under-fitted one, an over-fitted one and a just right one. Thus, it is true that there is a trade-off between the current data and the generalisation ability with unseen data. Data pre-processing (DP) is one of the crucial and time-consuming activities within Knowledge Discovery in Databases (KDD) [4]. Likely, feature selection [12] is said to be one of the most widesprea approaches within DP. It pursues to pick up the most important features in order to simplify the model and predict more accurately.

This paper aims at analysing the situation obtained when CFS and CNS pick up only a very reduced number of attributes and to propose a strategy to deal with this drawback. The rest of this article is organized as follows: Sect. 2 describes some concepts about feature selection; Sect. 3 introduces our proposal; Sect. 4 shows the experimental results; finally, Sect. 5 states the concluding remarks.

2 Feature Selection

Feature selection (FS) is one of the possible approaches to reduce the dimensionality. It picks up among the original variables those that are better suited for the problem at hand [7]. There are different kinds of methods to contend with feature selection [11]. Filter methods are independent of the classifier, whereas wrapper methods use the inductive algorithm as the evaluation function. FS involves two phases: (a) to get a list of attributes according to an attribute evaluator and (b) to carry out a search on the initial list. All candidate lists would be evaluated using a measure evaluation and the best one will be returned. Correlation-based Feature Selection (CFS) [6] and Consistency-based search in feature selection (CNS) [3] are two powerful methods to deal with this problem. Both operate together with a search method such as Greedy Search or typically Best First. There are some desirable properties to be exhibited in the reduced feature space: (a) low dimensionality, (b) retention of sufficient information, (c) enhancement of separability in feature space for examples in different classes by removing effects du to noise attributes, and (d) comparability of features among examples in the same category [14]. A goal of feature selection is to avoid selecting too many or too few features than is necessary [16]. If too few features are selected, there is a good chance that information content in this set of features is low. On the other hand, if too many (irrelevant) features are selected, the effects due to noise present in (most real-world) data may overshadow the information present. Hence, this is a trade-off which must be addressed by any feature selection method.

3 Proposal

The current paper sheds light on how to deal with problems where feature subset selection procedures choose a small number of attributes or even those cases when the solution reached for both procedures is exactly the same. As representative feature subset selectors we have singled out CFS and CNS. CFS is continually used from the developing time by Mark A. Hall up to now and currently have more than two thousands of citations and around fifty from the current year. To give some examples, one of the most recent work has been produced in the context of activity monitoring [13]. CNS is a popular, but not so much as the aforesaid, feature subset selector falling also into filter category with around seven hundred citations. CNS has been claimed that sometimes chooses a small number of features that may not be enough to provide the classifier an appropriate performance [20].

The first step is to collect the number of selected attributes for the aforementioned two methods for problems with a number of features less or equal than one hundred and a number of instances not higher than fifty thousand. As usually in data analytics field, the data preparation techniques operate on the training set and the test set remains unchanged and is evaluated by the first time after the classifier training. We have experimented with problems from the UCI (University of California at Irvine) repository [2] as well as some classical problems from the literature on Machine Learning [17] or Artificial Intelligence [15]. We have analysed around a hundred of problems. We have divided the problems into four outstanding scenarios as follows: (a) Only one selected attribute, (b) Only two attributes are singled out by CFS, (c) Only two picked attributes by CNS and (d) Same solution is achieved by CFS and CNS. As a second step, we also evaluated the classification performance with a previous data preparation from the original data with a feature ranking method such as ReliefF [9] which is a very strong method; later, it was extended to a very detailed work in [10] and by coincidence is now marking the twentieth birthday. ReliefF requires a threshold that may not be appropriately tuned according to the no free-lunch Theorem [23]. We go further and we propose Leave-k-out ReliefF or ReliefF$(-k)$, to shorten, which is the application of ReliefF with the dropping of k attributes. The value of k is set depending on the number of original attributes. In this study, for problems with lower than 10 attributes we have experimented with values of 1,2,3 for k. Next, for data sets with a number of attributes in the range (10,20] k takes the values 2,3,4. And finally, for problems containing more attributes, that is between 21 and 30, k is configured with the values 3,4,5. We have not defined more sets of values for k because we did not come across with any problem not fulfilling the condition in the four scenarios explained earlier.

4 Experimental Results

Table 1 describes the data sets utilised. Some of them are classical from Supervised Machine Learning and the remaining ones are publicly available in the UCI (University of California at Irvine) repository [2]. They come from real-world applications of different fields such as Medicine, Public Health, Botanic or Biology. The problems have been sorted by the number of attributes selected by CFS and CNS in ascending order. The size of the problems meeting the conditions of the four described scenarios ranges from around twenty to more than three thousand. The number of features varies between four and twenty nine, whereas the number of classes is between two and more than twenty. The missing values have been replaced in the case of nominal variables by the mode or, when concerning continuous variables, by the mean, bearing in mind the full data set.

We have evaluated the original and the reduced data sets with four classifiers such as C4.5 [18], 1NN [1], PART [5] and SVM [21]. We have used the implementations provided by WEKA tool [22] with default parameters that are those

Table 1. Summary of the data sets used and selected features for each feature subset selector

Data set	Instances			Classes	Features		
	Total	Train	Test		Original	CFS	CNS
Liver	345	259	86	2	6	1	–
Post-op	90	67	23	3	20	1	1
Lenses	24	18	6	3	6	1	2
Golf	28	20	8	2	4	2	2
Iris	150	111	39	3	4	2	2
Hypo	3772	2829	943	4	29	2	6
Breast	286	215	71	2	15	4	2
Smoking	2855	2141	714	3	13	5	2
Primary-tumor	339	251	88	21	23	5	5
Ecoli	336	251	85	8	7	6	6
Yeast	1484	1112	372	10	8	7	7
Average	882.6	661.3	221.4	5.5	12.3	3.3	3.5
Max	3772	2829	943	21	29	7	7

suggested by the own authors of the algorithms. We have chosen CFS and CNS as representative feature subset selection methods, because they are based on different kind of measures, have few parameters and have provided a good performance inside the supervised machine learning area. Often, BestFirst search is the preferable option by the researchers for both FSS algorithms. CFS is likely the most used FSS in data mining. CNS is also powerful, however the quantity of published works is more reduced [19].

4.1 Scenario I

This subsection copes with problems where only one attribute is obtained after the CFS. Table 2 shows the results of the experiments in this first scenario. According to the results, it seems that CFS is not able to capture the outstanding relationships between the variables and the class label. It is especially dramatic in the case of 1NN. CNS picks up from 0 through 2 attributes depending on the problem at hand. With two features the situation has been relieved a bit but in the remaining cases the performance is not very promising in general terms and sometimes very poor. Particularly, in Post-op the number of attributes selected by CFS and CNS is only a 5% of the original set which is extremely reduced and only C4.5 classifier is able to get a better performance. Generally speaking, ReliefF(–k) get more stable results keeping at least a half of the features in data sets with less than ten attributes or up to a 80% of the characteristic space for a problem with twenty features. As a general idea, 1NN does not operate very well with only one selected feature.

Table 2. Scenario I: Accuracy test results

Data set	Classifier	FULL	CFS	CNS	ReliefF (–1)	ReliefF (–2)	ReliefF (–3)	ReliefF (–4)
Liver	*C*4.5	68.60	58.14		59.46	68.60	66.28	
	1*NN*	61.63	39.53		58.30	54.65	56.98	
	PART	61.63	58.14		59.46	68.60	66.28	
	SVM	58.14	58.14		57.92	58.14	58.14	
Ind. average		62.50	53.49		58.78	62.50	61.92	
Post − op	*C*4.5	52.17	56.52	56.52		52.17	52.17	52.17
	1*NN*	56.52	4.35	4.35		56.52	56.52	56.52
	PART	65.22	56.52	56.52		56.52	56.52	56.52
	SVM	56.52	56.52	56.52		56.52	56.52	56.52
Ind. average		57.61	43.48	43.48		55.43	55.43	55.43
Lenses	*C*4.5	66.67	50.00	66.67	66.67	66.67	66.67	
	1*NN*	16.67	50.00	66.67	50.00	83.33	66.67	
	PART	66.67	50.00	66.67	66.67	66.67	66.67	
	SVM	66.67	50.00	66.67	66.67	83.33	66.67	
Ind. average		54.17	50.00	66.67	62.50	75.00	66.67	
Global averages	*C*4.5	62.48	54.89			62.48	61.71	
	1*NN*	44.94	31.29			64.83	60.06	
	PART	64.50	54.89			63.93	63.16	
	SVM	60.44	54.89			66.00	60.44	
Partial averages	*C*4.5			61.59	63.06			52.17
	1*NN*			35.51	54.15			56.52
	PART			61.59	63.06			56.52
	SVM			61.59	62.29			56.52
ReliefF(−2) *is a reasonable alternative to CFS and CNS*								

The value of k affects clearly to the performance but a trade-off value may be two, thus ReliefF(–2) is a quite reasonable solution to the problem showed by CFS and CNS. PART decreased the accuracy with Post-op after feature selection from any of the two poles, that is when only one attribute is selected or even almost all the features are retained. ReliefF(–k) shows a very flat behaviour in Post-op because the results remained constant.

4.2 Scenario II

We move on to the outlook where CFS only singles out two attributes which results are shown in Table 3. CNS picks from two to six attributes. Iris is a classical problem in the literature; according to the results at least two out of four attributes are required to generalised with a high accuracy. Golf is a data set rooted from the first studies in the field of Artificial Intelligence and the removal of one attribute with ReliefF(–1) keeps the original results. If more

Table 3. Scenario II: Accuracy test results

Data set	Classifier	FULL	CFS	CNS	ReliefF (–1)	ReliefF (–2)	ReliefF (–3)	ReliefF (–4)	ReliefF (–5)
Golf	*C4.5*	62.50	62.50	62.50	62.50	62.50	62.50		
	1NN	75.00	50.00	50.00	75.00	50.00	37.50		
	PART	62.50	62.50	62.50	62.50	37.50	62.50		
	SVM	37.50	37.50	37.50	37.50	50.00	25.00		
Ind. average		59.38	53.13	53.13	59.38	50.00	46.88		
Iris	*C4.5*	94.87	94.87	94.87	94.87	94.87	94.87		
	1NN	94.87	94.87	94.87	94.87	94.87	89.74		
	PART	94.87	94.87	94.87	94.87	94.87	94.87		
	SVM	94.87	94.87	94.87	94.87	94.87	94.87		
Ind. average		94.87	94.87	94.87	94.87	94.87	93.59		
Hypo	*C4.5*	99.15	96.92	98.94			99.26	99.26	98.94
	1NN	90.99	96.50	94.27			90.88	90.99	90.88
	PART	98.83	96.92	98.94			98.83	98.83	98.73
	SVM	93.85	93.32	93.43			93.85	93.85	93.85
Ind. average		95.70	95.92	96.39			95.71	95.73	95.60
Global averages	*C4.5*	85.51	84.77	85.44			85.54		
	1NN	86.95	80.46	79.72			72.71		
	PART	85.40	84.77	85.44			85.40		
	SVM	75.41	75.23	75.27			71.24		
Partial averages	*C4.5*			85.44	78.69	78.69		99.26	
	1NN			79.72	84.94	72.44		90.99	
	PART			85.44	78.69	66.19		98.83	
	SVM			75.27	66.19	72.44		93.85	

ReliefF(−2) or ReliefF(−1) are suitable for low − dimensionality problems
ReliefF(−k) with k = 3, 4, 5 is a good way for Hypo (initially 29 features)

than one attribute is discarded, the situation could be acceptable for k = 2 in general terms, but for k = 3 only in PART happens improvements compared with k = 2. Hypo is a problem which original results are better than those with the reduced sets. Although, the situation is appropriate with CNS. ReliefF(–k) is a good approach and the original results are even overcome. The conclusion is very simple and could us to assert that two attributes or three are at least necessary for small problems. Hypo is a problem which exhibits a strong classifier dependency but ReliefF(–4) is a good way.

4.3 Scenario III

This subsection depicts in Table 4 the results of those problems where only two attributes where retained by CNS. The situation achieved is very limited because with only two attributes out of more than twelve is not very easy to get good results for a classifier in general terms. CFS picked up from four through five attributes. Roughly speaking, the classification performance is very pretty with the exception of 1NN in Smoking data set. The removal of around a 20% of attributes let to recover more or less similar results that the full feature space.

Table 4. Scenario III: Accuracy test results

Data set	Classifier	FULL	CFS	CNS	ReliefF (–2)	ReliefF (–3)	ReliefF (–4)
Breast	*C*4.5	70.42	69.01	69.01	70.42	70.42	70.42
	1*NN*	64.79	70.42	70.42	67.61	69.01	69.01
	PART	69.01	71.83	69.01	64.79	63.38	67.61
	SVM	64.79	66.20	64.79	66.20	66.20	64.79
Ind. average		67.25	69.37	68.31	67.25	67.25	67.96
Smoking	*C*4.5	68.63	69.47	69.47	67.65	69.47	69.47
	1*NN*	54.76	38.52	5.60	56.86	50.28	50.14
	PART	61.48	67.36	68.77	62.75	61.76	66.11
	SVM	69.47	69.47	69.47	69.47	69.47	69.47
Ind. average		63.59	61.20	53.33	64.18	62.75	63.80
Global averages	*C*4.5	69.53	69.24	69.24	69.03	69.95	69.95
	1*NN*	59.78	54.47	38.01	62.23	59.65	59.58
	PART	65.25	69.60	68.89	63.77	62.57	66.86
	SVM	67.13	67.83	67.13	67.83	67.83	67.13
ReliefF(−*k*) *with k near to* 3 *exhibits a good performance*							

Table 5. Scenario IV: Accuracy test results

Data set	Classifier	FULL	CFS/CNS	ReliefF (–1)	ReliefF (–2)	ReliefF (–3)	ReliefF (–4)	ReliefF (–5)
Ecoli	*C*4.5	82.35	82.35	82.35	80.00	77.65		
	1*NN*	82.35	82.35	82.35	83.53	72.94		
	PART	80.00	80.00	80.00	78.82	76.47		
	SVM	83.53	83.53	83.53	83.53	77.65		
Ind. average		82.06	82.06	82.06	81.47	76.18		
Primary − tumor	*C*4.5	45.46	42.05			40.91	40.91	43.18
	1*NN*	36.36	30.68			36.36	37.50	37.50
	PART	43.18	42.05			40.91	39.77	38.64
	SVM	47.72	42.05			48.86	48.86	47.73
Ind. average		43.18	39.20			41.76	41.76	41.76
Yeast	*C*4.5	54.84	54.03	54.57	52.68	53.49		
	1*NN*	48.39	49.46	49.46	49.46	48.92		
	PART	56.72	54.30	55.65	55.91	54.30		
	SVM	55.91	54.84	54.57	54.30	53.76		
		53.97	53.16	53.56	53.09	52.62		
Global averages	*C*4.5	60.88	59.48			57.35		
	1*NN*	55.70	54.17			52.74		
	PART	59.97	58.78			57.23		
	SVM	62.39	60.14			60.09		
Partial averages	*C*4.5			68.46	66.34		40.91	43.18
	1*NN*			65.91	66.50		37.50	37.50
	PART			67.82	67.37		39.77	38.64
	SVM			69.05	68.92		48.86	47.73
ReliefF(−2) *is convenient for low − dimensionality data sets*								
ReliefF(−*k*) *with k* = 3, 4, 5 *is a good way for Primary − tumor* (23 *classes*)								

4.4 Scenario IV

Table 5 represents an overview where CFS and CNS pick up exactly the same features. The problems are now very handicapped because the number of classes is between eight and ten and the number of attributes is very close to the possible class labels. The situation is very delimited because the margin to discard attributes is not big because CFS and CNS have removed a single attribute for low-dimensionality problems and these are data sets are very complex especially to high number of classes. In the aforesaid scenario two attributes may be removed with ReliefF(–k) keeping a good performance. Contrarily, in Primary-tumor only five features are selected with CFS. Interestingly, ReliefF(–k) could discard safely at least three attributes that is at least a reduction close to a 15%.

5 Conclusions

A new feature ranking method called Leave-k-out ReliefF, also named ReliefF($-k$), was introduced. It was proposed as the alternative methodology when CFS or CNS only pick a very reduced number of attributes that could be 1 or 2 for any of these methods or even the same attributes are singled out by both methods. The recommendations depend on the number of original attributes and according to the test-bed are as follows. For problems with fewer than 10 attributes 1 or two attributes could be discarded safely. Finally, for problems with more features around 3 attributes could be removed from the input space.

Acknowledgments. This work has been partially subsidized by TIN2014-55894-C2-R project of the Spanish Inter-Ministerial Commission of Science and Technology (MICYT), FEDER funds and the P11-TIC-7528 project of the "Junta de Andalucía" (Spain).

References

1. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Mach. Learn. **6**(1), 37–66 (1991)
2. Bache, K., Lichman, M.: UCI machine learning repository (2013)
3. Dash, M., Liu, H.: Consistency-based search in feature selection. Artif. Intell. **151**(1), 155–176 (2003)
4. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Mag. **17**(3), 37 (1996)
5. Frank, E., Witten, I.H.: Generating accurate rule sets without global optimization (1998)
6. Hall, M.A.: Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato (1999)
7. Han, J., Pei, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier, Amsterdam (2011)
8. Kelleher, J.D., Mac Namee, B., D'Arcy, A.: Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press, Cambridge (2015)

9. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Bergadano, F., Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994). doi:10.1007/3-540-57868-4_57
10. Kononenko, I., Šimec, E., Robnik-Šikonja, M.: Overcoming the myopia of inductive learning algorithms with RELIEFF. Appl. Intell. **7**(1), 39–55 (1997)
11. Langley, P.: Selection of Relevant Features in Machine Learning. Defense Technical Information Center, Fort Belvoir (1994)
12. Liu, H., Motoda, H.: Computational Methods of Feature Selection. CRC Press, Boca Raton (2007)
13. Martinez-Mozos, O., Sandulescu, V., Andrews, S., Ellis, D., Bellotto, N., Dobrescu, R., Ferrandez, J.M.: Stress detection using wearable physiological and sociometric sensors. Int. J. Neural Syst. **27**(02), 1650041 (2017)
14. Meisel, W.S.: Computer-oriented approaches to pattern recognition. Technical report. DTIC Document (1972)
15. Michalski, R.S., Carbonell, J.G., Mitchell, T.M.: Machine Learning: An Artificial Intelligence Approach. Springer Science & Business Media, Berlin (2013)
16. Piramuthu, S.: Evaluating feature selection methods for learning in data mining applications. Eur. J. Oper. Res. **156**(2), 483–494 (2004)
17. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
18. Quinlan, J.R.: C4. 5: Programming for Machine Learning. Morgan Kauffmann, Burlington (1993)
19. Tallón-Ballesteros, A.J., Hervás-Martínez, C., Riquelme, J.C., Ruiz, R.: Improving the accuracy of a two-stage algorithm in evolutionary product unit neural networks for classification by means of feature selection. In: Ferrández, J.M., Álvarez Sánchez, J.R., Paz, F., Toledo, F.J. (eds.) IWINAC 2011. LNCS, vol. 6687, pp. 381–390. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21326-7_41
20. Tallón-Ballesteros, A.J., Riquelme, J.C., Ruiz, R.: Merging subsets of attributes to improve a hybrid consistency-based filter: a case of study in product unit neural networks. Connect. Sci. **28**(3), 242–257 (2016)
21. Vapnik, V.N.: The nature of Statistical Learning Theory. Springer, New York (1995)
22. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., Mining, D.: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington (2016)
23. Wolpert, D.H.: The supervised learning no-free-lunch theorems. In: Roy, R., Köppen, M., Ovaska, S., Furuhashi, T., Hoffmann, F. (eds.) Soft Computing and Industry, pp. 25–42. Springer, London (2002)