Feature selection and interpretable feature transformation: a preliminary study on feature engineering for classification algorithms

Antonio J. Tallón-Ballesteros [1[0000-0002-9699-1894]], Milan Tuba [2[0000-0003-3794-3056]], Bing Xue [3[0000-0002-4865-8026]] and Takako Hashimoto [4]

[1] University of Seville, Seville, Spain
[2] Singidunum University, Belgrade, Serbia
[3] Victoria University of Wellington, Wellington, New Zealand
[4] Chiba University of Commerce, Konodai Ichikawa City (Chiba), Japan
atallon@us.es

Abstract. This paper explores the limitation of consistency-based measures in the context of feature selection. These kinds of filters are not very widespread in large-dimensionality problems. Typically, the number of selected of attributes is very small and the ability to do right predictions is a drawback. The principal contribution of this work is the introduction of a new approach within feature engineering to create new attributes after the feature selection stage. The experimentation on multi-class problems with a feature space in the order of tens of thousands shed light on that some improvements took place with the new proposal. As a final insight, some new relationships were discovered due to the combined application of feature selection and feature transformation. Additionally, a new measure for classification problems which relates the number of features and the number of classes or labels is also proposed.

Keywords: Classification, Feature Engineering, Feature Selection, Data Mining, Feature Discovery, Feature transformation

1 Introduction

Classification studies problems where every object is categorised with a label represented in a discrete domain and therefore the number of values is limited [1]. The goal of classification is to predict the output variable value of unseen data given instances with values for the input and output variable(s). We have chosen two problems from Bioinformatics where the number of features is higher than twelve thousands and the number of classes is greater or equal than seven. As we are proposing a refinement of a previous approach, we have picked up two data sets with an error rate higher than a twenty percent and in the worst scenario in the sense that we are only considering the problems in a not very fruitful initial situation. According the reported results in a previous paper [2], the feature selection based on a stochastic search procedure such

as scatter search considering a consistency measure may be a fast approach although at the same time the number of selected characteristics is very low and the potential results are worse than with the application of classical correlation measures. This paper utilises as a baseline feature subset the resulting one with a scatter-search metaheuristic with a consistency-based measure which is only applied in the training set. Data preparation is conducted by means of feature selection in order to decrease the number of input feature space. Then, the features are augmented

Our goal is to improve the averaged test accuracy and the Cohen's kappa. The rest of this paper is organized as follows: Section 2 remembers some concepts about feature engineering and data mining; Section 3 introduces the proposal; Section 4 describes the experimental design; Section 5 depicts the results; finally, Section 6 states the conclusions.

2 Feature engineering and data mining

CRoss Industry Standard Process for Data Mining (CRISP-DM) proposes a framework to conduct data mining projects in an independent way of both the industry sector and the technology used [3]. This paper makes use of two visual tools to cover tasks from Data Analytics (DA) within Data Engineering (DE) [4] such as Waikato Environment for Knowledge Analysis (WEKA) [5] and RapidMiner [6]. Feature engineering (FE) plays an outstanding role in DA; it encompasses many fields such as feature transformation, feature generation, feature selection, feature analysis and many others. Machine learning algorithms cannot operate without data. As stated by G. Dong and H. Liu in the newly book published in 2018, "little can be achieved if there are few features to represent the underlying data objects, and the quality of the results of those algorithms largely depends on the quality of the available features" [7]. This idea has many connections with an earlier study from 2017 which put on the table the real situation that as times the feature selection may get a very small number of attributes which may not be enough to achieve reliable predictions [8]. Nowadays the data generation is going faster and faster and a data preparation task is an important step [9]. This paper puts emphasis on the feature perspective. Feature transformation (FT) is a process through which a new set of features is created [10]. Feature selection (FS) is a crucial task to conduct the training of classifiers with a reduced number of inputs and, at the same time, the outcome predictions could be more reliable. Nonetheless, a final characteristic-space with very small number of features constitutes an undesirable situation. There are some works accounting for the limited quality of the predictions and even the indiscernibility of the classes in scenarios with a very simple feature subset [8].

The bibliographical review and our previous experience motivate us to start the current research. A simple approach may be to create new features from the full feature space and then applying feature selection. Alternatively, we have typically applied feature selection and then we have applied supervised machine learning algorithms from the classification scope. Some papers aimed at increasing the final feature

space combining solutions from different types of feature selection approaches [11] or even iterating in a particular feature selection method more than once [12].

Thinking abstractly about all the aforementioned ideas and considering that an interpretable feature space is convenient we opt to deepen FT which is a way to augment the feature space combining groups or subgroups of features. FT could be conducted from the starting point with the whole feature space. On the first contribution in this field, FT was thought as an initial step and the application of FS may retrieve only the important properties of a study [10]. Our particular view is to establish a trade-off between FS and FT to be applied in contexts with continuous features. It is very well-known that for most of the problems in the nature, correlation-based relationships [13] are more common that consistency-based ones [14, 15, 16].

Basically, classifiers could be grouped in decision trees, neural networks, ruled based and classifiers based on the *k* nearest neighbours (*k*NN). We have chosen two reference algorithms such as *k*NN and Support Vector Machine (SVM) due to its good performance and their high consideration within data mining community.

3 Proposal

The proposal is based on the application of two kinds of FE procedures. The former carries out feature selection. The latter takes as input the selected features to conduct a feature transformation combining every different pair of selected attributes and applying to both operands a binary arithmetical operator (e.g. sum) to get one new attribute for every possible pair. Then, the selected features and the transformed features are merged to establish a new final characteristic space. As the proposal framework applies an operator which keeps the interpretative perspective on the data, we have called it Feature Selection and Interpretable Feature Transformation (FS-IFT). Fig. 1 depicts the approach proposed. There are no requirements for the current proposal although we would like to remark some important facts: i) any kind of feature selection approach may be conducted in the step 1 which is related with the initial feature selection. It means that feature ranking or feature subset selection are perfectly applicable, ii) the number of selected features must not be very high especially whether afterwards is going to be conducted in the feature transformation a pair combination of selected features; as a recommendation, no more than 25 attributes. For instance, a number of features around twenty is going to generate a number of new attributes which is close to two hundred.

This paper also introduces a new measure for classification to relate the number of features and the number of classes to characterise better classification problems. The Feature-to-class ratio (FtoC ratio) is a normalised ratio which represents how many features are approximately for every class label. The aforementioned ratio is of especial interest for feature engineering since the number of attributes after the data preprocessing is going to be altered in any way. It aids to analyse in a more detailed way the final outcome of the data preparation procedure. FtoC ratio is defined as follows:

$$\textit{Feature-to-class-ratio} = \textit{No. features}/\textit{No. classes} \quad (1)$$

FtoC gives a measurement between the dimensionality and the complexity in terms of number of different labels for a problem at hand. The number of patterns is also another interesting value to analyse; nonetheless we are working with Bioinformatics problems and their number is from almost one hundred to typically no more than five hundreds.

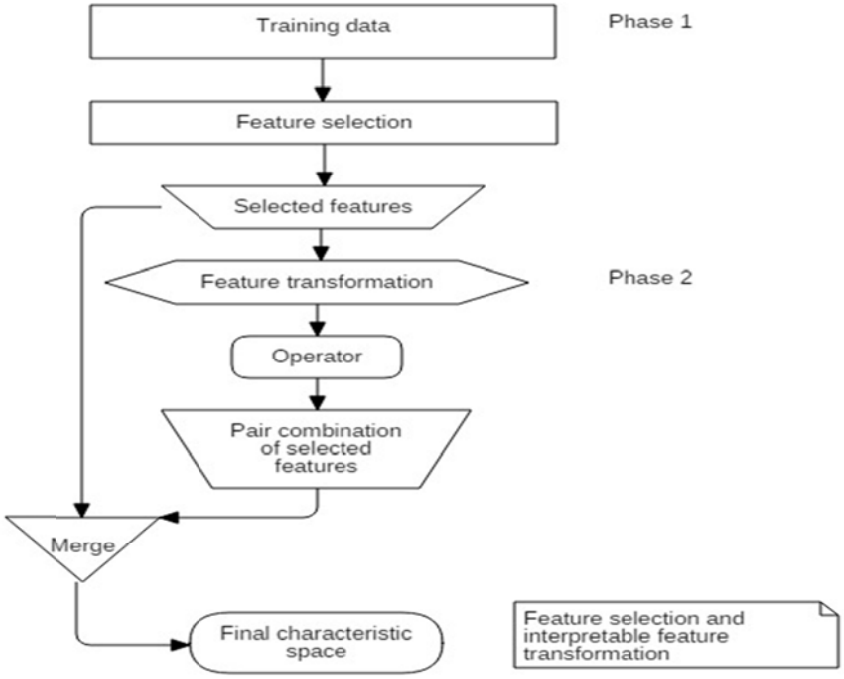

Fig. 1. Proposal framework within feature engineering.

4 Experimentation

Two challenging multiple class classification problems from Bioinformatics have been used to assess the proposal. Global Cancer Map (GCM) is a very difficult problem from the context of Bioinformatics which started to be studied in the beginning of the current century although its interest, even now, is very wide especially to the prevalence of very common diseases. GCM is populated with 190 tumor objects that have one out of the 14 possible labels such as Breast, Lung, Bladder to cite some of them in a very large dimensional space. Subtypes of Acute Lymphoblastic Leukemia (SALL) deals with leukemia and contains 327 samples distributed in the six known leukemia subtypes (T-cell, E2APBX1, TEL-AML1, MLL, BCR-ABL and hyperdiploid) and an extra category for those instances which do not belong to none of the featured subtypes. The order of magnitude of the number of attributes is exactly the same as SALL which is the order of tens of thousands. Additionally, the Feature-to-

class ratio is similar and its values are in-between one and two thousands of attributes with a distance to the cut-off point of an exact thousand around two hundred. Table 1 illustrates the most representative properties of the test bed. The experimental design follows a stratified hold-out cross validation procedure with three and one quarters for the training and testing sets, respectively.

Table 1. Summary of problems.

Problem	Patterns	Features	Classes	Feature-to-class ratio	Distribution of classes
GCM	190	16063	14	1147.4	10(3),11(8),20(1),22(1),30(1)
SALL	327	12558	7	1794.0	15(1),20(1),27(1),43(1),64(1),79(2)
Average	258.50	14310.50	10.50	1470.68	

Table 2 details the setting of the FE methods within the FS-IFT approach. Specifically, the proposal has been configured as follows: the first phase carries out feature subset selection based on a consistency-based measure with a scatter search to guide the exploration; the second phase takes as input the reduced training subset according to the previous step and creates new attributes with the arithmetical combination of the feature values that are merged to the selected attributes in the first step to create a new feature space to be used as the training set to train the classifier. As the baseline approach for feature selection, Scatter Search under a CoNsistency-based feature Selection (SS-CNS) [2] has been considered and the parameter that we have changed is the operator included in the second stage, hereinafter we refer to a concrete configuration of the proposal as FS-IFT(SS-CNS,op) where op could take as values the sum (+) or the multiplication (*). These operators are appropriate since the test bed consists of problems with continuous features. The feature engineering methods have been conducted for five different seeds to smooth the results and get reliable results due to the stochastic nature of the first procedure and the indirect stochasticity that is inherited into the second phase from first one. The combination of a stochastic step and a non-stochastic one further than all that is incorporated in the first stage has been applied within feature selection but not in the context of FE merging to subtypes of methods [17]. Table 3 reports the average number of attributes included in the feature space in the starting point, after the first phase of FS-IFT and at the end of FS-IFT which is also related to the number of classes to get an overview about what is convenient for a concrete problem.

5 Results

We have compared the baseline results (SS-CNS) to the new proposal in two scenarios such as FS-IFT(SS-CNS,+) and FS-IFT(SS-CNS,*). *k*NN (with k=1) and SVM classifiers are applied to two different databases in the contexts aforementioned in Sect. 4. Both classifiers have been assessed with two performance measures that are averaged with five seeds: accuracy and Cohen's kappa (CK). Tables 4 and 5 show the test results for *k*NN and SVM including the mean and the Standard Deviation (SD).

The sum operator helps to overcome the baseline results for GCM with kNN concerning the mean accuracy and/or the CK. There is one tie in the mean accuracy, although the SD is a bit slower what indicates more homogeneous solutions. For SALL, the improvements only happen with the multiplication operator which may reveal that the interaction of measures is more profitable to study Leukemia cases. Clearly, the new proposal is very convenient for SVM as there are gains in both scenarios for both problems. It is especially remarkable that with sum operator the improvements are very outstanding. As a final breakthrough, a Feature-to-class ratio in the environment of 1 (average values between 0.9 and 1.3) is not very promising to get accurate predictions according to the results. The experimentation has shown that with average values close to 6 the prediction ability of the classifier is considerably higher. Additionally, as one reviewer suggested, we have tried an extra feature selection procedure after FS-IFT. We have applied CoNsistency-based feature selection [15] and the number of selected attributes is even lower than the initial values after Phase 1.

Table 2. Feature engineering methods within Feature Selection and Interpretable Feature Transformation (FS-IFT).

Phase and name	Descriptive property	Value/s
Phase 1: Feature selection	Input file	Training set
	Input	Full feature space
	Output	Subset of features
	Attribute evaluation measure	Consistency
	Type of search	Scatter search
	Population size	250
Phase 2: Feature transformation	Input file	Reduced training set
	Input	Reduced feature space
	Output	Augmented reduced feature space
	Number of attributes to act as operands	2
	Operator	Sum(+), Multiplication(*)
	Way to obtain new attributes	All possible combinations excluding self-combinations

Table 3. Average number of attributes and Feature-to-class ratio in every phase of FS-IFT.

Problem	Features	Classes	Initial feature-to-class ratio	Selected features				Feature-to-class ratio			
				FS-IFT (Phase 1)		FS-IFT (Both phases)		FS-IFT (Phase 1)		FS-IFT (Both phases)	
				Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD
GCM	16063	14	1147.4	12.6	3.65	91.0	49.76	0.9	0.26	**6.5**	3.55
SALL	12558	7	1794.0	8.8	0.84	43.4	7.89	1.3	0.12	**6.2**	1.13

6 Conclusions

This paper presented a framework within feature engineering including two kinds of methods: the first step reduces the dimensionality via feature selection and the second one expands the feature space by means of feature transformation. The proposed reported very competitive results with remarkable improvements in one or two assessment measures. Additive and multiplicative relations are very noticeable. In kNN there are some cases where the sum is better, whereas in others the multiplication is better, while in SVM the sum is substantially better in all cases. Concretely, for GCM the classifier kNN obtained an improvement of 2.61 in accuracy with the + operator and in SALL it is obtained an improvement of 0.93 with * operator. With the classifier SVM, in GCM it is obtained an improvement of 2.61 as well as in SALL a gain of 2.44 with is also accompanied with very relevant overcoming in CK evaluation measure. Additionally, the feature-to-class ratio must be similar to the number of classes as happens in the second most difficult case as SALL is or close to the half as in the most challenging problem (GCM). Though the current proposal has been focused on two complex problems from Bioinformatics, we strongly believe that the proposed approach could be extensively used for other classification problems involving an important number of attributes and multiple classes and even for other classifiers. Moreover, this contribution may represent an important push for promoting the analysis of low dimensionality spaces that may be achieved after feature subset selection especially with consistency measures or sometimes with correlation metrics.

Table 4. *K*NN classifier: Test results.

Problem	Feature Engineering procedures											
	SS-CNS				FS-IFT(SS-CNS,+)				FS-IFT(SS-CNS,*)			
	Accuracy		CK		Accuracy		CK		Accuracy		CK	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
GCM	41.30	3.64	0.3630	0.0390	**43.91**	3.19	**0.3910**	0.0340	41.30	3.37	0.3640	0.0350
SALL	80.98	3.24	0.7666	0.0038	80.73	3.03	0.7630	0.0360	**81.91**	3.40	**0.7770**	0.0410
Average	61.14		0.5648		62.32		0.5770		61.61		0.5705	

Table 5. SVM classifier: Test results.

Problem	Feature Engineering procedures											
	SS-CNS				FS-IFT(SS-CNS,+)				FS-IFT(SS-CNS,*)			
	Accuracy		CK		Accuracy		CK		Accuracy		CK	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
GCM	30.43	5.67	0.2200	0.0750	**33.04**	5.40	**0.2690**	0.0570	31.30	6.24	0.2350	0.0770
SALL	81.46	3.96	0.7670	0.0490	**83.90**	4.11	**0.7999**	0.0520	82.93	4.29	0.7870	0.0550
Average	55.95		0.4935		58.47		0.5345		57.12		0.5110	55.95

Acknowledgment. This work has been partially subsidised by TIN2014-55894-C2-R, TIN2017-88209-C2-R (Spanish Inter-Ministerial Commission of Science and Technology (MICYT)), P11-TIC-7528 projects ("Junta de Andalucía" (Spain)) and FEDER funds. It has also been supported by the Ministry of Education, Science and Technological Development of Republic of Serbia, Grant no. III-44006.

References

1. Alpaydin, E.: *Introduction to machine learning*. MIT press (2014).
2. Tallón-Ballesteros, A.J, Ibiza-Granados, A.: Simplifying pattern recognition problems via a scatter search algorithm. *International Journal for Computational Methods in Engineering Science and Mechanics*, *17*(5-6), 315-321 (2016).
3. Wirth, R., Hipp, J.: CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39) (2000).
4. Cho, S.B., Tallón-Ballesteros, A.J.: Visual Tools to Lecture Data Analytics and Engineering. In *International Work-Conference on the Interplay Between Natural and Artificial Computation* (pp. 551-558). Springer, Cham (2017).
5. Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I.H., Trigg, L.: Weka-a machine learning workbench for data mining. In *Data mining and knowledge discovery handbook* (pp. 1269-1277). Springer, Boston, MA (2009).
6. Akthar, F., Hahne, C.: Rapidminer 5 operator reference. *Rapid-I GmbH*, *50*, 65 (2012).
7. Dong, G., Liu, H.: Feature Engineering for Machine Learning and Data Analytics. CRC Press (2018).
8. Tallón-Ballesteros, A.J., Riquelme, J.C.: Low dimensionality or same subsets as a result of feature selection: an in-depth roadmap. In *International Work-Conference on the Interplay Between Natural and Artificial Computation* (pp. 531-539). Springer, Cham (2017).
9. Tallón-Ballesteros, A.J., Li, K. (Eds.).: *Fuzzy Systems and Data Mining III: Proceedings of FSDM 2017* (Vol. 299). Ios Press (2017).
10. Liu, H., Motoda, H.: Feature transformation and subset selection. *IEEE Intelligent Systems*, (2), 26-28 (1998).
11. 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. *Connection Science*, *28*(3), 242-257 (2016).
12. Tallón-Ballesteros, A.J., Correia, L., Xue, B.: Featuring the Attributes in Supervised Machine Learning. In *International Conference on Hybrid Artificial Intelligence Systems* (pp. 350-362). Springer, Cham (2018).
13. Hall, M.A.: Correlation-based feature selection for machine learning (1999).
14. Shin, K., Kuboyama, T., Hashimoto, T., Shepard, D.: sCwc/sLcc: Highly Scalable Feature Selection Algorithms. *Information*, *8*(4), 159 (2017).
15. Dash, M., Liu, H., Motoda, H.: Consistency based feature selection. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 98-109). Springer, Berlin, Heidelberg (2000).
16. Arauzo-Azofra, A., Benitez, J.M., Castro, J.L.: Consistency measures for feature selection. *Journal of Intelligent Information Systems*, *30*(3), 273-292 (2008).
17. Tallón-Ballesteros, A.J., Correia, L., Cho, S.B.: Stochastic and Non-Stochastic Feature Selection. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 592-598). Springer, Cham (2017).