Taylor & Francis
Taylor & Francis Group

Journal of Decision Systems

ISSN: 1246-0125 (Print) 2116-7052 (Online) Journal homepage: http://www.tandfonline.com/loi/tjds20

Data-driven decision-making in classification algorithm selection

Dijana Oreški & Nina Begičević Ređep

To cite this article: Dijana Oreški & Nina Begičević Ređep (2018): Data-driven decision-making in classification algorithm selection, Journal of Decision Systems, DOI: 10.1080/12460125.2018.1468168

To link to this article: https://doi.org/10.1080/12460125.2018.1468168

Published online: 21 May 2018.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=tjds20

JOURNAL OF DECISION SYSTEMS, 2018
https://doi.org/10.1080/12460125.2018.1468168

Taylor & Francis
Taylor & Francis Group

Check for updates

Data-driven decision-making in classification algorithm selection

Dijana Oreški and Nina Begičević Ređep

Faculty of Organization and Informatics, University of Zagreb, Varazdin, Croatia

ABSTRACT

The selection of the appropriate classification algorithm for a given data-set is an important and complex issue, full of research challenges. In this paper, we present a developed meta-analysis-based framework to improve decision-making in the selection of classification algorithms based on data-set characteristics. We study the effectiveness of our proposed framework with 32 data-sets. Three classification algorithms – neural networks, decision trees, and k-nearest neighbours – were trained and applied to data-sets with different characteristics, aiming to review the performance of algorithms in the presence of noise in the data, the interaction between features, as well as a small or a large ratio between the number of instances and the number of features. Our results show that feature noise is the most important predictor of the decision regarding the choice of the classification algorithm, and data-driven classification is found to be useful in this scenario.

ARTICLE HISTORY
Received 11 January 2018
Accepted 11 April 2018

KEYWORDS
Data characteristics; data-driven classification; CRISP DM; Decision-making; meta-learning

1. Introduction

The volume of data stored by organisations grows exponentially every day. To gain competitive advantage in a competing environment, organisations are forced to extract knowledge from data. All these factors have prompted the need for intelligent data analysis methodologies, which could discover useful knowledge from data (Misra & Dehuri, 2007). The term *data mining* refers to the process of knowledge discovery in data. While a large number of data mining methods have been established for numerous problems, many challenges remain to be overcome. Currently, the basic question for the data mining community is how to select the appropriate technique. Our paper addresses the problem of data mining technique selection based on the data characteristics. Specifically, we focus on the data mining classification task. Our central argument is the need to employ a framework for the classification algorithm selection to improve classification performance. To do so, extensive experiments involving three types of classification algorithms (statistical, machine learning and neural computing approach) five types of data-set characterisations (number of features, number of instances, data sparsity, correlation of features, feature noise), and all possible numbers of the data-sets were conducted on the 32 publicly available data-sets. Thus, 480 analyses were performed.

CONTACT Dijana Oreški dijana.oreski@foi.hr

© 2018 Informa UK Limited, trading as Taylor & Francis Group

The main aim of our research is to identify the data-set characteristics associated with classification algorithm performance, using three classifiers, five data-set characteristics and thirty-two data-sets with different characteristics. Our work's contributions can therefore be interpreted as follows. We investigated feature noise and data sparsity whose impact on classification algorithm performance so far has not been investigated. Furthermore, we have applied decision tree in evaluation of the results. This approach shown to be very useful. Empirical analysis findings revealed feature noise and data sparsity as most important data-set characteristics in predicting classification algorithm performance.

This paper is organised in the following structure. Section 2 provides the literature review on the given topic. Section 3 explains the methodology through the CRoss-Industry Standard Process for Data Mining (CRISP DM) standard model applied in this research. Comparative analyses on classification algorithms (techniques) are presented in Section 4. The final section concludes our study.

2. Literature review

Many researchers have focused on developing new classification techniques, giving rise to the need to determine which technique to use in a given situation. According to the 'no free lunch' theory, no best technique exists for all situations (Peng, Wang, Kou, & Shi, 2011). It is imperative to find an appropriate technique; therefore, the main aim of our research is to define which technique to use in a specific situation. Existing approaches use the trial-and-error method, and there is a lack of systematic research concerning which classification technique should be used on a particular data-set, based on the characteristics of the data-set. Data-set characteristics are key in determining the classification algorithm's performance (Kwon & Sim, 2013). Previous papers have proven that the choice of the 'best' classification algorithm depends on the given data-set (Ali & Smith, 2006; Bernado-Mansilla & Ho, 2005; Chen & Shyu, 2011; Dessì & Pes, 2015; Kiang, 2003; Kwon & Sim, 2013; Smith, Woo, Ciesielski, & Ibrahim, 2002; Song, Wang, & Wang, 2012). Kiang (2003) points out that data characteristics affect the performance of classification methods. Bernado'-Mansilla and Ho (2005) tried to determine the appropriate algorithm for a classification problem. Ali and Smith's (2006) study suggests that 'a more useful strategy is to gain an understanding of the data-set characteristics that enable different learning algorithms to perform well, and to use this knowledge to assist learning algorithm selection based on the characteristics of the data-set'. Chen and Shyu (2011) claim that the correlations among data-set characteristics affect algorithm performance, but few studies have analysed the influence of data-set characteristics on classification algorithm performance.

In this paper, we aim to investigate how data-set characteristics affect the performance of classification algorithm techniques. We consider accuracy as a measure of algorithm performance. A decision tree (DT) is used to determine the relationships among the techniques, using the data-set characteristics as the independent variables and classification accuracy as the dependent variable. As such, our research makes significant contributions in following aspects: (i) we investigate data-set characteristics that were not yet explored in these terms (feature noise), (ii) we includes algorithms that were not yet explored in these terms (kNN), (iii) we have measured data-set characteristics impact on the classification algorithms performance by means of decision tree – which we see as the most significant novelty of our approach.

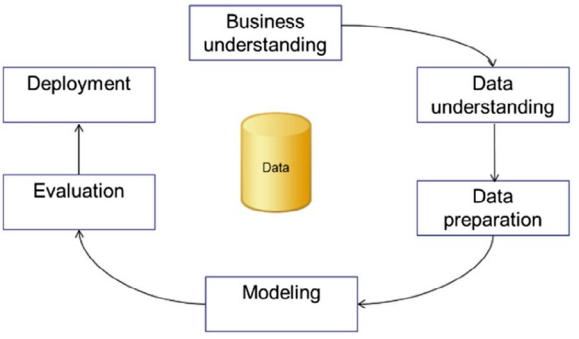

Figure 1. CRISP DM framework.

Table 1. Data-set characteristics (Van der Walt, 2008).

Data-set characteristic	Abbreviation	Sources
Standard measures		Michie, Spiegelhalter, and Taylor (1994); Sohn (1999); Song et al. (2012); Van der Walt (2008)
Number of features (dimensionality)	*d*	
Number of instances	*N*	
Data sparsity measures		Anand and Bharadwaj (2011); Van der Walt (2008)
Data sparsity	*DS*	
Statistical measures		Michie et al. (1994); Sohn (1999); Van der Walt (2008)
Correlation of features	*p*	
Noise measures		Gamberger, Lavrac, and Dzeroski (2000); López et al. (2013); Van der Walt (2008); Zhu and Wu (2004)
Feature noise	*ID2*	

3. Research methodology

The data mining process must be reliable and repeatable by people with little data mining background. With this aim, several standards have been developed, as follows: Knowledge Discovery in Data (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), Sample-Explore-Modify-Model-Assess (SEMMA) (Olson & Delen, 2008) and CRISP DM (Wirth & Hipp, 2000). CRISP DM is the most frequently used (Azevedo & Santos, 2008), which we apply here to develop the data-driven framework for the classification algorithm selection. CRISP DM consists of six steps, as presented in Figure 1.

The first step focuses on understanding the objectives and defining the data mining problem and the goals of the analysis (Wirth & Hipp, 2000). In the first phase of the framework, the data-set characteristics that are relevant to the classification task are identified. Van der Walt (2008) examined the relationship between data-set characteristics and classifier performance and developed data measures to define this relationship. These measures are grouped into the following categories: standard, data sparsity, statistical and noise measures (Table 1). From the 11 characteristics in Van der Walt's research (Van der Walt, 2008), we have extracted five, used in at least one other previous study. These are number of instances, number of features, data sparsity, correlation and feature noise.

The second step of the framework, data understanding, involves exploring the data and verifying its quality (Azevedo & Santos, 2008). Data understanding starts with an initial data collection and proceeds with activities to become familiar with the data. In this step, the data-sets were collected. To find data-sets with different characteristics, we used three publicly available data repositories, as follows: the Sociology Data Set Server of Saint Joseph's

University in Philadelphia (2017), StatLib (2017), and the UCI Machine Learning Repository (2017). Thirty-two data-sets were extracted to cover all combinations of different characteristics – five characteristics with two states each (small and large).

The third step is data preparation, covering all activities required to prepare the data-set for modelling. Here, we calculated the values of the characteristics of each data-set. The fourth step, modelling, consists of building and assessing the models. Prior, modelling technique should be selected. Three modelling techniques (DT, neural network [NN] and k-nearest neighbours [kNN]) were selected and applied, and their parameters were calibrated to optimal values (Wirth & Hipp, 2000). These three techniques were chosen for two reasons. First, they represented different approaches to learning: statistical (kNN), machine learning (DT), and NN. Second, they were used in previous comparative analyses of classification techniques, as follows: DT (Ali & Smith, 2006; Kiang, 2003; Song et al., 2012), NN (Ali & Smith, 2006; Brazdil, Gama, & Henery, 1994; Kiang, 2003), and kNN (Brazdil et al., 1994; Kiang, 2003; López, Fernández, García, Palade, & Herrera, 2013). The fifth step is the evaluation, which answers the question of how well the model performs on the test data and interprets the model. In the evaluation phase, DT modelling was performed to gain insights into the interdependencies among the different approaches to classification and data characteristics. Sixth, the deployment step determines how the results should be utilised and who needs to use them (Azevedo & Santos, 2008). The guidelines based on the rules established by the DT were extracted here.

4. Research results

With the aim of investigating whether the data-set characteristics would affect the classification algorithm performance, the DT analysis was conducted. Its purpose was to compare the results of different classifiers on the same set of data-sets. The C5.0 algorithm was employed in the fourth step of our research to extract the rules that showed the associations between the data-set characteristics (input variables) and the classification algorithm performance (output variables). The results showed that DTs provided specific and detailed depictions of the associations between the data-set characteristics and the performance of techniques. Figure 2 presents the DT model. The target attribute in the root node was the classification algorithm. The first-level attribute was found to be the feature noise; the second-level attributes were the number of features and data sparsity.

Based on the results of the fourth step of our research rules are extracted in Table 2. We suggest using the DT when dealing with a high level of large feature noise. This approach can handle noise very well. Feature noise measures how many features are not contributing significantly to classification. Quantification of feature noise is calculated as explained in (Oreski, Oreski, & Klicek, 2017). Based on the previous results, some guidelines are suggested. When dealing with a high level of feature noise in the data, the DT is an excellent choice. However, the performance of these algorithms deteriorates in the presence of a small number of features. When dealing with data-sets with a low level of feature noise, the kNN and NNs present promising behaviour. The kNN performs adequately with low levels of data sparsity, whereas with higher levels of sparsity, NNs' accuracy improves.

We conducted a sensitivity analysis in the sixth step of our research to detect the importance of the characteristics through the column contribution. The column contribution

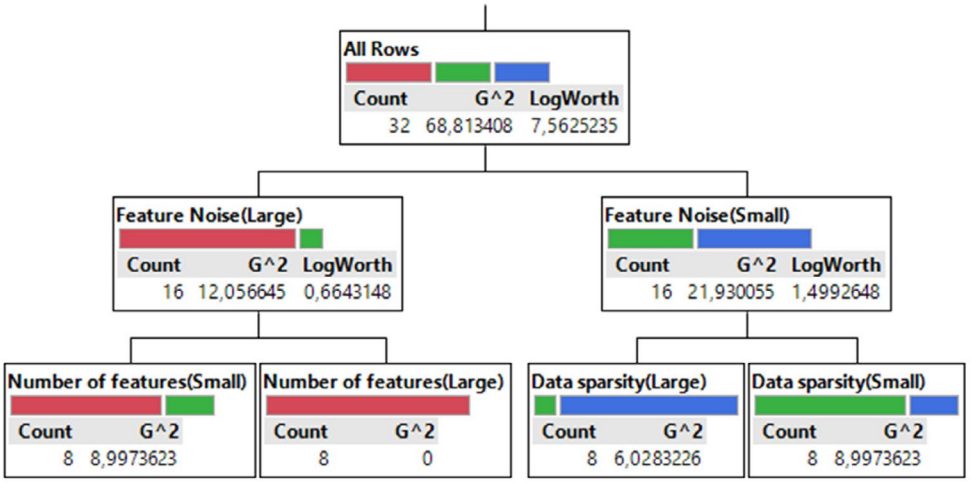

Figure 2. Decision tree of dependencies between data characteristics and classification.

Table 2. Rules of dependencies between data characteristics and classification algorithm.

Leaf label	Decision tree	kNN	NN
Feature noise (large) & number of features (small)	0.7199	0.2518	0.0283
Feature noise (large) & number of features (large)	0.9421	0.0296	0.0283
Feature noise (small) & data sparsity (large)	0.0440	0.1440	0.8120
Feature noise (small) & data sparsity (small)	0.0440	0.6996	0.2564

Table 3. Contribution of data-set characteristics to algorithm selection.

Term	G^2	Portion
Feature noise	34.826708	0.7775
Data sparsity	6.90436974	0.1541
Number of features	3.05928285	0.0683
Number of instances	0	0.0000
Correlation	0	0.0000

indicates the measure with which each data characteristic provides the information to predict the value of the classification technique. Table 3 illustrates the relative importance of the data-set characteristics.

Most of the data-set characteristics are significant, except the number of instances and the correlation. All characteristics indicate that they are related to the classification algorithm performance. In the DT model, the coefficients of independent variables, each represented as a portion, indicate that the unit change in the data characteristics contributes to the change in the classification algorithm performance. Nonetheless, comparing the portion scores of the data characteristics provides a general insight into how the DT model interprets

Table 4. Validation of decision tree model.

Measure	Training
Generalised R^2	0.8260
Mean-Log p	0.4209
RMSE	0.3642
Mean Abs Dev	0.2845
Misclassification Rate	0.1563
N	32

Table 5. *T*-test.

Classification algorithm	*p*-value
kNN	0.024
NN	0.015
DT	–

the importance of each input aspect. In this study, feature noise is the most important data characteristic for consideration when deciding which classification algorithm to apply on a certain data-set.

Overall, the DT models for the five data-set characteristics achieve satisfactory reliability (Table 4), with R^2 of 0.8260 and the classification rate of 0.8437.

Based on pairwise *t*-test shown in Table 5, on average, the overall accuracy of the DT algorithm is better than the overall average accuracy of all the other algorithms and the difference is statistically significant at 95% confidence level in favour of the DT.

5. Conclusion

Our paper contributes to existing research on data mining by demonstrating how classification, guided by knowledge about data-set characteristics, enhances data analysis. We have demonstrated the relationship between data-set characteristics and classification technique performance on the data-set with particular characteristics, so we assume that if the characteristics of other data-sets are similar, then the performances of the classification techniques on these data-sets are similar as well. Thus, our results could serve as guidelines for classification technique selection. Data-sets with high feature noise should use decision tree in classification. Data-sets with small feature noise and small data sparsity should apply kNN approach in classification. Data-sets with small feature noise and large data sparsity sholud employ neural networks. Given the outstanding model-fitting results, the DT finds feature noise to be the most significant aspect, while the least important ones are the number of instances and the correlation among the features. Such findings reiterate prior studies' results that some classification algorithms are extremely sensitive to noise, and feature noise is the most vital attribute in differentiating a particular algorithm from others. Conclusively, it is suggested that examining the data characteristics is the best way to choose the most suitable classification technique. Our research has a few limitations. Only a specific group of data-set characteristics is included. Thus, the group of characteristics could be extended to imbalanced ratio, information-theoretic measures and statistical measures. In our future research, we will explore the relationships between a larger number of data-set characteristics and increase the number of classification algorithms to include a broader spectrum.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Dijana Oreški http://orcid.org/0000-0002-3820-0126

References

Ali, S., & Smith, K. A. (2006). On learning algorithm selection for classification. *Applied Soft Computing, 6*(2), 119–138.

Anand, D., & Bharadwaj, K. K. (2011). Utilizing various sparsity measures for enhancing accuracy of collaborative recommender systems based on local and global similarities. *Expert Systems with Applications, 38*(5), 5101–5109.

Azevedo, A. I. R. L., & Santos, M. F. (2008). Kdd, semma crisp-dm: A parallel overview. IADS-DM. Retrieved March 25, 2017, from https://www.openaire.eu/search/publication?articleId=od_____2595::e3a84b91c9fc5298ec939f0d80353d81

Bernado-Mansilla, E., & Ho, T. K. (2005). Domain of competence of XCS classifier system in complexity measurement space. *IEEE Transactions on Evolutionary Computation, 9*(1), 82–104.

Brazdil, P., Gama, J., & Henery, B. (1994, April). Characterizing the applicability of classification algorithms using meta-level learning. In European Conference on Machine Learning (pp. 83–102). Berlin, Heidelberg: Springer.

Chen, C., & Shyu, M. L. (2011, August). Clustering-based binary-class classification for imbalanced data sets. In Information Reuse and Integration (IRI), 2011 IEEE International Conference on (pp. 384–389). Las Vegas, NV: IEEE.

Dessì, N., & Pes, B. (2015). Similarity of feature selection methods: An empirical study across data intensive classification tasks. *Expert Systems with Applications, 42*(10), 4632–4642.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 2011*(11), 27–34.

Gamberger, D., Lavrac, N., & Dzeroski, S. (2000). Noise detection and elimination in data preprocessing: Experiments in medical domains. *Applied Artificial Intelligence, 14*(2), 205–223.

Kiang, M. Y. (2003). A comparative assessment of classification methods. *Decision Support Systems, 35*(4), 441–454.

Kwon, O., & Sim, J. M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications, 40*(5), 1847–1857.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113–141.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). *Machine learning, neural and statistical classification*. New York.

Misra, B. B., & Dehuri, S. (2007). Functional link artificial neural network for classification task in data mining. *Journal of Computer Science*, *3*(12), 948–955.

Olson, D. L., & Delen, D. (2008). Advanced data mining techniques. Berlin: Springer Science & Business Media.

Oreski, D., Oreski, S., & Klicek, B. (2017). Effects of dataset characteristics on the performance of feature selection techniques. *Applied Soft Computing, 52*, 109–119.

Peng, Y., Wang, G., Kou, G., & Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction. *Applied Soft Computing, 11*(2), 2906–2915.

Smith, K. A., Woo, F., Ciesielski, V., & Ibrahim, R. (2002). Matching data mining algorithm suitability to data characteristics using a self-organizing map. In *Hybrid information systems* (pp. 169–179). Physica: Heidelberg.

Sohn, S. Y. (1999). Meta analysis of classification algorithms for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(11), 1137–1144.

Sociology Data Set Server of Saint Joseph's University in Philadelphia. Retrieved November 10, 2017, from http://sociology-data.sju.edu/

Song, Q., Wang, G., & Wang, C. (2012). Automatic recommendation of classification algorithms based on data set characteristics. *Pattern Recognition, 45*(7), 2672–2689.

StatLib – Carnegie Mellon University. Retrieved November 10, 2017, from http://lib.stat.cmu.edu/

UCI Machine Learning Repository. Retrieved November 10, 2017, from http://archive.ics.uci.edu/ml/datasets.html

Van der Walt, C. M. (2008). *Data measures that characterise classification problems* (Doctoral dissertation). Pretoria.

Wirth, R., & Hipp, J. (2000). 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*. New York, NY

Zhu, X., & Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review, 22*(3), 177–210.