Framework of Intelligent System for Machine Learning Algorithm Selection in Social Sciences

Dijana Oreški*
University of Zagreb, Faculty of Organization and Informatics, Varazdin, Pavlinska 2, Croatia

* Corresponding author. Tel.: +385 42 390 860; email: dijana.oreski@foi.hr
Manuscript submitted January 17, 2021; accepted May 11, 2021.
doi: 10.17706/jsw.17.1.21-28

Abstract: The ability to generate data has never been as powerful as today when three quintile bytes of data are generated daily. In the field of machine learning, a large number of algorithms have been developed, which can be used for intelligent data analysis and to solve prediction and descriptive problems in different domains. Developed algorithms have different effects on different problems. If one algorithm works better on one dataset, the same algorithm may work worse on another data set. The reason is that each dataset has different features in terms of local and global characteristics. It is therefore imperative to know intrinsic algorithms behavior on different types of datasets and choose the right algorithm for the problem solving. To address this problem, this paper gives scientific contribution in meta learning field by proposing framework for identifying the specific characteristics of datasets in two domains of social sciences: education and business and develops meta models based on: ranking algorithms, calculating correlation of ranks, developing a multi-criteria model, two-component index and prediction based on machine learning algorithms. Each of the meta models serve as the basis for the development of intelligent system version. Application of such framework should include a comparative analysis of a large number of machine learning algorithms on a large number of datasets from social sciences.

Key words: Data features, intelligent system, machine learning, meta learning.

1. Introduction

Over the last few years we have witnessed a huge number of machine learning algorithms applications in a broad spectrum of domains. They have a crucial role in harnessing the power of the vast amount of data we produce daily in the digital age. The application of these algorithms and the process of developing quality predictive and descriptive models is complex, iterative and time-consuming because it involves comparing a large number of algorithms and adjusting their parameters. Therefore, there is a need to automate the selection of algorithms for models development. In this paper, we are dealing with definition of framework which would serve as a basis for intelligent system development in the field of social sciences. The paper is structured as follows. Section two gives in depth overview of relevant papers published in the field. Section 3 explains methodology and section 4 gives suggestion of framework. Section 5 concludes the paper and gives guidelines for further research.

2. Related Work

Many research papers were focused on developing modelling algorithms and a large number of algorithms have been developed, so there is a need to determine which algorithm is best to use in a specific

situation, in a particular domain, at a particular dataset. According to the "no free lunch" theory, no best technique exists for all situations (Peng et al., 2011). It is therefore imperative to selectively employ appropriate modelling techniques. Existing approaches generally use a "trial and error" basis and there is a lack of systematic research concerning which modelling technique should be used on a particular dataset, based on the characteristics of this dataset. With the of learning from previous experience, on data driven way, meta learning field was developed. In context of machine learning, meta learning is process of learning from previous experience gained by application of different algorithms on datasets of different characteristics (Brazdil et.al., 2008). Several approaches to meta learning were developed: learning based on properties of task, learning based on prior evaluation and learning based on prior trained models (Vanschoren, 2018). Previous research in the domain of classification techniques suggests that dataset characteristics considerably impact the performance of the machine learning algorithms and proves that the choice of the "best" classification algorithm is dependent on the given dataset (Chen and Shyu, 2011., Dessì and Pes, 2012., Kwon and Sim, 2013). Little recent research has focused on these issues. Kiang suggests that data characteristics considerably impact the performance of classification techniques (Kiang, 2003). Bernadoı ´ −Mansilla and Ho (Bernado´−Mansilla and Ho, 2005) have developed metrics to evaluate the ability of a classification techniques to characterize different datasets, and use those metrics to determine the appropriate technique for a new classification problem. Ali and Smith's study on the classification techniques selection problem (Ali and Smith, 2006) confirms that it is necessary to understand dataset characteristics to assist in learning algorithm selection. Chen and Shyu claim that the correlation between the characteristics of a dataset affect the performance of techniques (Chen and Shyu, 2011). The selection of a proper techniques for a specific classification problem is very difficult, as the choice of the techniques to use depends on the chosen dataset (Song and Wang, 2012). Due to the large amount of data and availability of techniques for analysis, and given to resource constraints, research in the past year has been especially directed to automatic selection of techniques (Franklin, 2018). Therefore, we are listing a series of recent papers. Porto et. al. (2018) provide a comprehensive experimental comparison and propose a meta-learning solution designed for automatic selection and recommendation. On the same track is research of Ali and his associates. They associate dataset characteristics with selection of classification techniques but focus only on different decision tree algorithms and do not include other approaches to classification nor do they take into account dimensionality reduction or descriptive modelling techniques (Ali et al., 2018). Wu and Lu (2018) argue that researchers so far have, to a certain extent in some domains, developed automated methods for selecting algorithms based on data characteristics, however, such approach is not applicable to data sets from other domains. Zhang et al. (2019) in the research published a few days ago focused on one dataset characteristics and by extensive comparison of techniques tried to determine which is the best for certain characteristic. However, they do not take into account the domain specificity. Research by Sivakumara et al. (2018) also explored this topic by comparing techniques in the medical domain for early detection of cancer. Their comparative analysis of supervised learning techniques on different datasets resulted in a proposal for classification technique for that domain. Lorena et al. (2018) point out the need for data driven selection of dimensionality reduction techniques and analyzed characteristics that would give guidelines for selection of dimensionality reduction techniques. Authors emphasized the importance of exploring the domain characteristics of datasets. Based on this review, in this paper we are focusing of domain of social sciences.

3. Data and Methods

Why social sciences domain? Everyday social activities are influenced by digital systems which generate huge amounts of data when interacting with users. These data are used in social science research. The availability of data sources and the development of algorithms and tools for data collection and analysis, improved social science research since recording of phenomena that were previously unnoticed or even non-existent is enabled. Such data are social data because information systems are complex and dynamic, consisted of digital technology and social interaction. Whether it is a business system or a public database

containing information about entrepreneurial activity, collecting, processing and disseminating data through these systems reflects some form of social action. When analyzing data, the characteristics of the datasets must be taken into account. In this framework, we focus on two social science domains: education and business, combined forming a whole. From elementary to high school, and higher education, pupils and students' behavior data is generated based on which academic performance is predicted. The process of transforming raw data into structured datasets convenient for machine learning algorithm application is one of the most important tasks of predictive modelling in education (Gardner and Brooks, 2018). In addition to data preparation, Gardner and Brooks identified a number of challenges from a data analysis methodology perspective: the lack of model evaluation and holistic approach in selecting the best predictive models is just one of them. Furthermore, Li, Wang, and Wang (Li, Wang, and Wang, 2017) emphasize the need to consider features that have correlative, temporal, and fragmented properties in the development of predictive models in education since such aspects will increase interpretation and accuracy of predictions. This is motivation for our research. The business domain relies on education. Upon graduation, students enter the business world. In the domain of business, there are a number of valuable data sources. We highlight the GEM database, which is a valuable source of information on entrepreneurial activity. The GEM database is extensive and imposes specific challenges to data analysis. In the field of analytics in entrepreneurship, methodology aspect imposes several questions. Bergmann et al. argue that there is no systematic analysis of relations between dataset characteristics and methodological capabilities in the entrepreneurial domain (Bergmann, Mueller, & Schrettle, 2014), which imposes its specificities. For example, researchers of entrepreneurial activity should take into account they are dealing with "rare events" (one value of the dependent variable is more frequent than the other value - class imbalance problem) (Bergmann, Mueller, & Schrettle, 2014). Those are only some of open questions from previous research which served as the basis for setting up the goals of this research proposal.

4. Framework of Intelligent System

The process of intelligent data analysis consists of several phases defining to discover potentially useful information and knowledge from the "raw data". This process has been standardized and several standards have been developed so far. The most commonly used is the CRISP DM standard, which defines the following stages: domain understanding, data understanding, data preparation, modelling, model evaluation and usage. Within the standard, the modelling phase plays a crucial role, given the responsibility for extracting the patterns which will be used after the evaluation. Described protocol is applicable to problems of different domains. Each of these domains has its own characteristics that can qualify the data and specific data characteristic and features. The aim of framework presented here is to identify specific features of datasets from education and business domains and, at the end, develop an intelligent system for automatic recommendation of a modelling algorithm based on these features. The framework is presented in the figure below.

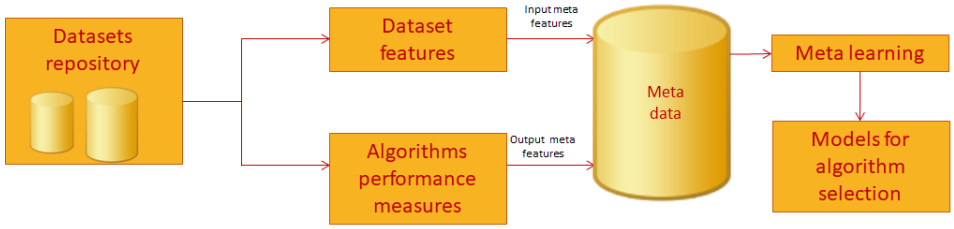

Fig. 1. Intelligent system framework.

The first phase, understanding the domain, focuses on the goals and requirements of the domain,

converting them to the definition of data mining problems, and developing a plan for achieving this goals. As part of first phase, the goals for each domain will be defined and domain experts will provide in depth explanation of domain requirements. The second phase, data understanding, begins with the initial data collection and continues with activities aimed at getting to know the data, detecting the quality of the data, and gaining first insight into dataset distributions. At this stage, data from different sources will be collected in education and business domain. Some of the data source in the domain of education are: data from e-learning systems, surveys conducted among students and publicly available repositories. In the domain of business, GEM (Global Entrepreneurship Monitor) database will be used, datasets from the World Bank Group Entrepreneurship Survey and publicly available repositories consisting of life quality data. Besides data collection, the basic activity of the second phase is data description, which involves examining the meta-features of the datasets. The relevant meta-features will be identified by a literature review. Up to 10 meta-features will be selected and a categorization of each feature will be performed. By selecting 10 features, each with a minimum of two categories, 1024 datasets need to be analyzed for each domain (a total of 2048 data sets) in order to capture all feature combinations. If there would be such combination of features, which could not be found on real dataset, such dataset will be generated. Models will be developed on each dataset using different approaches to machine learning. Supervised machine learning algorithms are categorized into four categories: probability-based algorithms, error-based algorithms, information-based algorithms, and similarity-based algorithms. The minimum number of machine learning algorithms to be included is four: which means that the minimum number of analyzes to be performed is 8192 analyzes. During the evaluation phase, the degree to which the model meets the business objectives is evaluated, and the measures of accuracy and reliability of the model (e.g., confusion matrix, precision, response, RSquare) will be used as for evaluation. Such results will be integrated into a meta dataset consisting of a meta-example, each meta-example consisting of meta features and a target attribute: ranking of the machine learning algorithm. Based on such meta dataset, meta models will be developed by means of different approaches: correlation and ranking, multi-criteria modelling, using supervised learning algorithms. Meta models will be evaluated and the best meta model will be built into the final version of the intelligent system. As a result, an intelligent system will be developed for automatic recommendation of the machine learning algorithm taking into account domain and characteristics of the dataset in that domain.

All obtained models will be interpreted with the respect to domain knowledge and success criteria. By doing so, such framework provides contributions in two directions: (i) methodologically, through the development of meta models and intelligent system, and (ii) domain-based, by discovering of new information, knowledge and patterns in each of the domains involved. Expected scientific contributions of application of such framework:

1) systematization of knowledge in the field of dataset meta-features relevant for social sciences,
2) repository of datasets in education and business domains with identified specific meta - features,
3) developed descriptive models in the fields of education and business;
4) developed predictive models dependent on the specific meta-features of education and business datasets,
5) developed and evaluated meta models for machine learning algorithm selection based on the meta - features of the datasets,
6) developed intelligent system for automatic selection of a machine learning algorithm, depending on the task and meta- features,
7) created guidelines for decision makers in the field of education and business.

In his merit, this meta learning framework defines how to develop reliable models in social sciences. In doing so, focus is on developing models capable of solving real-world problems in the field of education and

business. As shown in the literature review, intelligent data analysis is still under-applied in social sciences, and there are many open questions of how to use machine learning in particular domains of the social sciences. The framework proposed here is trying to give our contribution in this area. Besides scientific contributions, application of framework gives various social contributions. We believe that the real problems that will be addressed in intelligent system development are of great interest for society. Problems that will be addressed in the context of the education have a huge social impact. For example, the prediction of academic performance is a problem that interests both students and educational institutions, and the modelling of self-evaluation and the development of recommendation can contribute to the improvement of online teaching, which is taking a boom in all areas of society. On the other hand, entrepreneurial problems that are addressed in the intelligent system are of great importance at the social level due to the interest that exists on all related issues of quality of life. We believe that our results can achieve great contributions to the field of data science, but also help us address important social issues and that allow us to get in touch with the different agents to tackle socio-economic objectives and transfer the knowledge obtained through the development of software applications and recommendations. Moreover, there are multiple potential users of such system:

1) data scientists: The basis of the framework is the intelligent recommender system for data scientists - which techniques to use in data analysis, depending on the domain and characteristics of the data set. System is useful for experienced data scientists as well as domain experts with little knowledge of data analytics.
2) management of Higher Education Institutions, Teachers and Students: student predictive success models are important at all three levels because they explain the academic achievement and prevent a high level of student drop-outs from studies.
3) managers: Predictive models of entrepreneurial activity identify trends in entrepreneurship and give new insight into entrepreneurial intentions.
4) Governments and policy makers: through guidelines derived from business models.
5) In the future research we will implement this framework in order to develop intelligent system.

Acknowledgment

This work has been supported in part by Croatian Science Foundation under the project UIP-2020-02-6312."

References

[1] Ali, R., Khatak, A. M., Chow, F., & Lee, S. (2018, January). A case-based meta-learning and reasoning framework for classifiers selection. *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication*.

[2] Ali, S., & Smith, K. A. (2006). On learning algorithm selection for classification, *Appl. SoftComput*, *6(2)*, 119–138.

[3] Bergmann, H., Mueller, S., & Schrettle, T. (2014). The use of global entrepreneurship monitor data in academic research: A critical inventory and future potentials. *International Journal of Entrepreneurial Venturing*, *6(3)*, 242-276.

[4] Bernado, M., E., & Ho, T. K. (2005). Domain of competence of XCS classifier systemin complexity measurement space. *IEEE Trans. Evol. Comput.*, *9(1)*, 82–104.

[5] Bilalli, B., Abello, A., & Aluja, B. T. (2017). On the predictive power of metafeatures in OpenML. *International Journal of Applied Mathematics and Computer Science*, *27(4)*, 697–712.

[6] Brazdil, P. (2008). Christophe giraud carrier, carlos soares, and ricardo vilalta. *Metalearning:*

Applications to Data Mining. Springer Science & Business Media.

[7] Cantabella, M., Martínez-España, R., Ayuso, B., Yáñez, J. A., & Muñoz, A. (2019). Analysis of student behavior in learning management systems through a big data framework. *Future Generation Computer Systems*, *90*, 262-272.

[8] Cervone, D. *et al*. (2014). Predicting points and valuing decisions in real time with NBA optical tracking data. Retrieved from: http://www.sloansportsconference.com/wp-content/uploads/2014/02/2014_SSAC_Pointwise-Predicting-Points-and-Valuing-Decisions-in-Real-Time.pdf

[9] Chen, C., & Shyu, M. (2011). Clustering-based binary-class classification for imbalanced datasets, *Proceedings of 2011 IEEE International Conference on Information Reuse and Integration*.

[10] Cui, C., Hu, M., Weir, J. D., & Wu, T. (2016). A recommendation system for meta-modeling: A meta-learning based approach. *Expert Systems with Applications*, *46*, 33-44.

[11] Dessì, N., & Pes, B. (2015). Similarity of feature selection methods: An empirical study across data intensive classification tasks, *Expert Syst. Appl*., *42(10)*, 4632–4642.

[12] Divjak, B., & Oreski, D. (2009). Prediction of academic performance using discriminant analysis. *Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces* (pp. 225-230).

[13] Domo, data never sleeps. Retrieved from: https://www.domo.com/learn/data-never-sleeps-5?aid=ogsm072517_1&sf100871281=1

[14] Filipović, D., Balaban, I., & Oreški, D. (2018, January). Cluster analysis of students' activities from logs and their success in self-assessment tests. *Proceedings of the Central European Conference on Information and Intelligent Systems.*

[15] Franklin, F. (2018). Automatic selection of mapreduce machine learning algorithms: A model building approach.

[16] Gardner, J., & Brooks, C. (2018). Evaluating predictive models of student success: Closing the methodological gap.

[17] Hajdin, G., Hainš, V. V., & Oreški, D. (2018, January). The impact of teaching scenarios on student perception of teaching. *Proceedings of the in Edulearn18: 10th International Conference on Education and New Learning Technologies*.

[18] Kadoić, N., & Oreški, D. (2018, May). Analysis of student behavior and success based on logs in Moodle. *Proceedings of the 2018 41st International Convention on Information and Communication Technology*.

[19] Kedmenec, I., Oreški, D., Vuković, K., Postolov, K., & Jovanovski, K. (2017). Decision tree modelling for entrepreneurial intention. *Proceedings of the 11th MAC 2017*.

[20] Kiang, M. Y. (2003). A comparative assessment of classification methods. *Dec. SupportSyst*., 441–454.

[21] Kliček, B., Oreški, D., & Divjak, B. (2010). Determining individual learning strategies for students in higher education using neural networks. *International Journal of Arts and Sciences*, *3(18)*, 22-40.

[22] Kovač, R., & Oreški, D. (2018). Educational data driven decision making: Early identification of students at risk by means of machine learning. *Proceedings of the Central European Conference on Information and Intelligent Systems* (pp. 231-237).

[23] Kwon, O., Sim, J. M. (2013). Effects of dataset features on the performances of classification algorithms. *Expert Syst. Appl.*, *40*, 1847–1857.

[24] Li, X., Wang, T., & Wang, H. (2017, March). Exploring n-gram features in clickstream data for MOOC learning achievement prediction. *Proceedings of the International Conference on Database Systems for Advanced Applications* (pp. 328-339).

[25] Lorena, A. C., Garcia, L. P., Lehmann, J., Souto, M. C., & Ho, T. K. (2018). How complex is your

classification problem? A survey on measuring classification complexity.

[26] Maršić, K., & Oreški, D. (2016). Estimation and comparison of underground economy in Croatia and European union countries: Fuzzy logic approach. *Journal of Information and Organizational Sciences*, *40(1)*, 83-104.

[27] Obschonka, M. (2017). The quest for the entrepreneurial culture: psychological big data in entrepreneurship research. *Current Opinion in Behavioral Sciences*, *18*, 69-74.

[28] Oreski, D. (2013). Impact of data characteristics on feature selection techniques Performance. *Research Papers Faculty of Materials Science and Technology Slovak University of Technology*, 84-89.

[29] Oreški, D., & Begičević, R. N. (2018). Data-driven decision-making in classification algorithm selection. *Journal of Decision Systems*, 248-255.

[30] Oreški, D., & Kadoić, N. (2018, January). Analysis of ICT students' LMS engagement and sucess. *Proceedings of the 35th International Scientific Conference on Economic and Social Development–Sustainability from an Economic and Social Perspective*.

[31] Oreski, D., & Klicek, B. (2015). A novel feature selection techniques based on contrast set mining. *Proceedings of the 14th International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*.

[32] Oreški, D. & Konecki, M. (2015, January). Selection of classification algorithm using a meta learning approach based on data sets characteristics. *Proceedings of the Information Society–IS 2015*.

[33] Oreski, D., Kedmenec, I., & Klicek, B. (2016). Exploring capabilities of contrast mining application in SWOT analysis. *Proceedings of the 8th MAC 2016*, 210.

[34] Oreški, D., Konecki, M., & Milić, L. (2017, May). Estimating profile of successful IT student: Data mining approach. *Proceedings of the 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 723-727).

[35] Oreški, D., & Novosel, T. (2014). Comparison of feature selection techniques in knowledge discovery process. *TEM Journal*, *3(4)*, 285.

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

[37] Oreški, D., Pihir, I., & Kedmenec, I. (2016, January). The association between the national intellectual capital components and the quality of life. *Proceedings of the International Academic Multidisciplinary Research Conference 2016*.

[38] Oreski, D., Pihir, I., & Konecki, M. (2017). Crisp-DM process model in educational setting. *Economic and Social Development: Book of Proceedings*, 19-28.

[39] Oreški, D., Hajdin, G., & Klicek, B. (2016). Role of personal factors in academic success and dropout of IT students: Evidence from students and alumni. *TEM Journal*, *5(3)*, 371

[40] Oreski, S., Oreski, D., & Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert Systems with Applications*, *39(16)*, 12605-12617.

[41] Pasanisi, R., & Pasanisi, S. (2018). How to discover hidden knowledge according to different type data set: A guideline to apply the right hybrid information mining approach. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *9(4)*, 83-99.

[42] Peng, Y., Wang, G., Kou, G., & Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction, *Appl. Soft Comput.*, 2906–2915.

[43] Porto, F., Minku, L., Mendes, E., & Simao, A. (2018). A systematic study of cross-project defect prediction with meta-learning.

[44] Sivakumar, S., Nayak, S. R., Vidyanandini, S., Kumar, J. A., & Palai, G. (2018). An empirical study of

supervised learning methods for breast cancer diseases.
[45] Song, Q., Wang, G., & Wang, C. (2012). Automatic recommendation of classification algorithms based on dataset characteristics, *Pattern Recogn*., 2672–2689.
[46] Sunil, J. (2017). Why big data is the new game-changer in election. Retrieved from: https://www.linkedin.com/pulse/why-big-data-new-game-changer-elections-sunil-jose/
[47] Vanschoren, J. (2018). Meta-learning: A survey.
[48] Wu, M. S., & Lu, J. Y. (2018, July). Automated machine learning algorithm mining for classification problem. *Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition*.
[49] Zhang, X., Li, R., Zhang, B., Yang, Y., Guo, J., & Ji, X. (2019). An instance-based learning recommendation algorithm of imbalance handling methods. *Applied Mathematics and Computation*, *351*, 204-218.

Dijana Oreški was born in Varaždin in 1986. She has obtained her masters and PhD degree at the University of Zagreb, Faculty of Organization and informatics at the field of data science.

She works as an associate professor at the University of Zagreb, Faculty of Organization and informatics. Dijana Oreški is a member of various conferences program committees and journal editorial boards. She has received several awards for her scientific work.