ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Logarithmic division for green feature selection: an information-theoretic approach

Samuel Suárez-Marcote, Laura Morán-Fernández and Verónica Bolón-Canedo *

CITIC, Universidade da Coruña, A Coruña, Spain

Abstract. Feature selection is a popular preprocessing step to reduce the dimensionality of the data while preserving the important information. In this paper we propose an efficient and green feature selection method based on information theory, with the novelty of using the logarithmic division and resort to fixed-point precision. The results of experiments conducted on several datasets indicate the potential of our proposal, as it does not incur in significant information loss compared to the standard method, both in the features selected and in the subsequent classification step. This finding opens up possibilities for a new family of green feature selection methods, which would help to minimize energy consumption and carbon emissions.

1 Introduction

Artificial intelligence has achieved numerous advances in recent years. These are primarily due to the increase in computing power, the development of new algorithms and the emergence of big data. The latter has not only provided machine learning with a wealth of opportunities but has also significantly increased the dimensionality of data. Feature selection addresses this problem by selecting the relevant features from the data while discarding irrelevant or redundant ones [1].

Problems with high dimensionality are becoming commonplace in many environments, such as bioinformatics, where it is essential to identify the biomolecules that explain a particular phenotype [2]. This process enables a more straightforward and coherent explanation. However, feature selection often faces scalability problems, which limit its effectiveness. Due to the combinatorial nature of feature selection, having an extensive feature space can significantly impact its performance. Therefore, it is necessary to develop improvements to cope with the ever-increasing volume of data. Many of today's state-of-the-art improvements rely on high-performance computing, which requires significant computational resources and results in considerable energy consumption. However, the increasing number of Internet of Things (IoT) devices has led to the emergence of trends like Edge Computing [3], which aims to run algorithms at the end nodes of the network. Thus, Edge Computing can reduce data transfer and eliminate the need for supercomputers or cloud servers. Given the memory and energy limitations of IoT devices and the desire to obtain greener and more efficient algorithms,

*This work was supported by the Ministry of Science and Innovation of Spain (Grant PID2019-109238GB-C22 / AEI / 10.13039 / 501100011033) and together with "NextGenerationE"/PRTR (TED2021-130599A-I00) and by Xunta de Galicia (Grants ED431G 2019/01 and ED431C 2022/44).

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

some approaches in the field of neural networks aim to achieve this through quantization. As an efficient and underexplored alternative for feature selection, fixed-point representation is being proposed. For instance, a fixed-point 32-bit integer adder consumes nine times less energy than a single-precision floating-point adder [4]. In a previous work [5], we conducted a detailed study of the impact of precision loss on filter-type feature selection methods. The selection of these methods was based on their low computational cost. Specifically, the study narrowed down its focus to filter-type methods that employ metrics based on Information Theory due to its wide range of applications. In this paper, we propose a more realistic implementation of the chosen information theory-based methods. Thanks to this new approach, it becomes feasible to obtain more realistic measurements, which facilitates a more comprehensive analysis. As part of this, we also modify the computation of low-precision Mutual Information (MI) metric, which is utilized by several feature selection methods and thus encourages the implementation of new ones that will operate with low-precision.

Since modern algorithms involve complex arithmetic operations, optimizing these operations is critical. Division is very costly in terms of clock cycles and, hence, energy. In this context, logarithmic division is proposed as a promising approach to achieve more energy-efficient algorithms that can cope with the limitations of IoT devices. Although the conversion to the logarithmic scale presents some challenges, it is well documented that using the logarithmic number system can significantly improve time and energy consumption, particularly in applications where multiplications and divisions are common [6]. To date, the application of the logarithmic numeric system in feature selection has not been explored. Similarly to the use of low-precision, there is a growing tendency to utilize logarithmic division as a means of conserving time and memory in neural networks. This is why, we propose the use of logarithmic division for feature selection algorithms implemented in fixed-point to improve their efficiency.

In this work, we present the implementation of feature selection methods that employ fixed-point representation and logarithmic division to enhance energy, time and memory efficiency. A study will be conducted to analyze the loss of information resulting from the use of these approximations. Furthermore, the implementation of the low-precision Mutual Information metric enables its usage in other applications that rely on this metric.

2 Proposed methodology

To conduct this study, filter-type methods, which perform feature selection independently of the induction phase and have low computational cost, have been chosen. In addition, these methods reduce the dependence on the decided classifier. Within the group of filter-type methods, we have selected those based on the Mutual Information metric and narrowed it down to two: Mutual Information Maximization (MIM) and Joint Mutual Information Maximization (JMI) [7]. MIM has been selected for its simplicity, while JMI has been chosen for being a good trade-off between stability and accuracy [7].

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

In order to carry out the experimental process, the first step is to adapt the methods to work with fixed-point arithmetic. Fixed-point numbers are integers scaled by a constant factor, with a fixed number of digits in the fractional part. As stated earlier, fixed-point usage can lead to large energy savings as well as enable the operation of these algorithms in IoT devices and, therefore, their application to Edge Computing. Simultaneously, this approach improves security as it obviates the need to send all data to the cloud. For further study, we opted for a more realistic implementation based on the one provided by Meurant [8] for the Matlab tool, but some modifications have been introduced, including the addition of logarithmic division. Regarding the division, Mitchell's algorithm has been chosen, since it is known for its efficiency in terms of energy consumption, silicon area and time [9]. This, together with the simple implementation of this algorithm using fixed-point representation, makes it a promising option for obtaining even more energy-efficient feature selection methods.

Following the modification of the feature selection algorithms, we designed a two-phase experimental process, using the double-precision results as a baseline:

1. Comparison of the rankings obtained by the low-precision feature selection methods against the baseline. We use the True Positive Rate (TPR) to measure similarity. However, unlike the TPR used in classification, we define it as follows: $TPR = \frac{TP}{n}$, where TP represents the number of shared features between the two rankings and n denotes the ranking size.

2. Study of the results obtained in a subsequent classification step for the k-nearest neighbour classifier (k-NN), with a value of $k = 3$, a commonly used value in the literature. This classifier was chosen due to its minimal assumptions about the data. To estimate the error rate, we performed 3x5-fold cross-validation, which involved conducting both the feature selection and classification steps within a single cross-validation loop.

Based on the results of our previous work [5], which established that the size of the fractional part is the primary consideration when selecting representations for the experimental process, a fractional part size of 75% has been selected for the 32-bit, 16-bit, 8-bit and 4-bit representations. Consequently, this results in fractional part sizes of 24, 12, 6, and 3 bits. These experiments were conducted in a Matlab2022a and Weka 3.8.6 environment.

3 Experimental results

For the experimental process, four datasets have been selected, consisting of two datasets acquired via wearable devices and two microarrays. The microarray datasets are challenging because of their small sample size compared to a large number of features. Table 1 displays the primary characteristics of these datasets. It is important to note that as MIM and JMI are ranker methods, it is necessary to manually set a threshold for the number of relevant features. The value of this threshold (n) is directly related to the number of features in the

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

dataset. For *PhysicalActivity*, the threshold values of 5, 10, and 15 have been selected, while for the other datasets, the threshold values are 25, 50, and 75. Moreover, MI calculation is required to be performed on discrete data. Therefore, a discretization in 5 equal-width intervals has been carried out.

Dataset	#Samples	#Features	#Classes	Repository
physicalActivity (phys.)	6264	18	6	[10]
humanActivity (hum.)	7352	561	6	[10]
SRBCT* (SRB.)	83	2308	4	[11]
lung* (lung)	39	2880	2	[12]

Table 1: Main datasets properties. Star (*) identifies microarray-type datasets.

First of all, checking the similarity of the rankings obtained, Table 2 displays the mean value and deviation of TPR for each combination of thresholds and representations. Upon examining the baseline representations of the methods, we can see that MIM and JMI exhibit similar behaviour. The use of 32 bits resulted in identical rankings to the baseline and the TPR value decreased as the number of available bits was reduced. Additionally, rankings with smaller n values yielded lower TPR values since this metric is a ratio and making a mistake in the ranking has more impact on the similarity metric. Moving on to the results obtained by the versions that utilize logarithmic division, the behaviour was again similar for both methods. It stands out how the 32-bit and 16-bit representations obtained lower TPR values around 0.1 when compared to the previous versions without division. However, this trend did not hold for the 8-bit and 4-bit representations.

For the sake of brevity, and because in the previous experiments the highest threshold usually obtains the best results, in the following we will only report classification for this threshold, i.e., 15 features for the physicalActivity dataset and 75 features for the others. Table 3 shows the classification accuracy values obtained with k-NN after using MIM. It can be observed that the decrease in the similarity between rankings does not lead to an accuracy reduction. This is especially noticeable for the smaller 8-bit and 4-bit representations, where the obtained TPRs were low. In addition, the results obtained using logarithmic division versions are comparable to those obtained using the base implementation.

Results for the JMI method are presented in Table 4. Similar to the previous case, the differences in rankings do not directly imply a loss of information for the classification step. Once again, the logarithmic split versions do not exhibit significantly different results from those obtained by the baseline method, proving that using the proposed approach does not significantly impact the posterior classification performance while achieving benefits in energy consumption.

4 Conclusions

In this paper, we explore the use of fixed-point representation and logarithmic division to optimize filter-type feature selection methods in terms of efficiency

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Bits	n	Method			
		Base Method		Logarithmic Division	
		MIM	JMI	MIM	JMI
32	5 - 25	1.000±0.00	1.000±0.00	0.890±0.11	0.900±0.12
	10 - 50	1.000±0.00	1.000±0.00	0.940±0.04	0.950±0.05
	15 - 75	1.000±0.00	1.000±0.00	0.907±0.06	0.930±0.07
16	5 - 25	0.990±0.02	0.970±0.04	0.870±0.08	0.870±0.09
	10 - 50	0.995±0.01	0.960±0.05	0.930±0.03	0.910±0.06
	15 - 75	0.983±0.03	0.987±0.01	0.910±0.06	0.880±0.08
8	5 - 25	0.590±0.10	0.610±0.17	0.600±0.06	0.560±0.13
	10 - 50	0.695±0.13	0.705±0.09	0.700±0.07	0.675±0.15
	25 - 75	0.817±0.04	0.777±0.16	0.837±0.12	0.753±0.23
4	5 - 25	0.100±0.07	0.100±0.11	0.090±0.06	0.070±0.06
	10 - 50	0.230±0.11	0.210±0.20	0.255±0.12	0.195±0.15
	15 - 75	0.423±0.30	0.333±0.37	0.440±0.32	0.377±0.34

Table 2: Mean and standard deviation of TPR for the base and logarithmic versions of the MIM and JMI methods using the four low-precision representations (4, 8, 16 and 32 bits) for the different thresholds n of selected features.

Bits	Base Method				Logarithmic Division			
	SRB.	hum.	lung	phys.	SRB.	hum.	lung	phys.
64	98.413	85.727	72.685	70.402	100.00	85.564	74.352	70.062
32	98.413	85.727	72.685	70.402	99.603	85.863	75.185	69.923
16	98.413	85.714	72.685	70.259	100.00	85.895	74.352	70.035
8	96.825	95.598	66.759	71.121	97.163	94.895	72.685	69.955
4	87.619	87.908	63.519	68.487	87.897	94.709	68.333	68.093

Table 3: Average classification accuracy after applying MIM on low-precision approaches (4, 8, 16 and 32 bits) and the double-precision floating-point version (64 bits) for the largest n-top.

Bits	Base Method				Logarithmic Division			
	SRB.	hum.	lung	phys.	SRB.	hum.	lung	phys.
64	92.024	97.615	71.204	70.392	96.012	97.706	76.019	69.817
32	92.024	97.615	71.204	70.392	96.012	97.819	75.926	69.998
16	91.627	97.692	71.204	70.392	96.409	97.815	75.926	69.955
8	96.825	96.260	72.037	69.812	97.222	96.373	73.519	69.673
4	85.139	85.124	67.685	70.589	84.564	82.227	70.000	69.993

Table 4: Average classification accuracy after applying JMI on low-precision approaches (4, 8, 16 and 32 bits) and the double-precision floating-point version (64 bits) for the largest n-top.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

and energy consumption. We conducted experiments using two widely used information-theoretic filter methods, MIM and JMI, with different bit representations and threshold levels. Our findings indicate that using low-precision representations can potentially reduce energy consumption [4] without sacrificing information. Moreover, the use of logarithmic division allows for an additional degree of energy savings [6], which is especially useful for resource-constrained environments. Regarding the classification results, our study showed that the modified algorithms using fixed-point representation and logarithmic division performed comparably to the baseline implementations, even when the rankings obtained were less similar.

However, there is room for further improvement.To better optimize these savings, further studies are needed in real-world environments to properly evaluate time and memory consumption. Overall, these findings represent a promising step towards efficient and scalable machine learning applications.

References

[1] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

[2] H. Climente-González, C.-A. Azencott, S. Kaski, and M. Yamada. Block hsic lasso: model-free biomarker detection for ultra-high dimensional data. *Bioinformatics*, 35(14):i427–i435, 2019.

[3] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5):637–646, 2016.

[4] M. Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 10–14. IEEE, 2014.

[5] S. Suárez-Marcote, L. Morán-Fernández, and V. Bolón-Canedo. Less is more: Low-precision feature selection for wearables. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[6] B. Parhami. Computing with logarithmic number system arithmetic: Implementation methods and performance benefits. *Computers & Electrical Engineering*, 87:106800, 2020.

[7] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The journal of machine learning research*, 13:27–66, 2012.

[8] G. Meurant. Fixed point, floating point and posits. unpublished, `https://gerard-meurant.pagesperso-orange.fr/`, N.D.

[9] C. Subhasri, B. R. Jammu, L. Guna Sekhar Sai Harsha, N. Bodasingi, and V. R. Samoju. Hardware-efficient approximate logarithmic division with improved accuracy. *International Journal of Circuit Theory and Applications*, 49(1):128–141, 2021.

[10] Kaggle, Google LLC. Kaggle datasets. [Online; accessed April 2023]. `https://www.kaggle.com/datasets`.

[11] A. Statnikov, I. Tsamardinos, Y. Dosbayev, and C. F. Aliferis. Gems: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. *International journal of medical informatics*, 74(7-8):491–503, 2005.

[12] J. Li and H. Liu. Kent ridge bio-medical data set repository. *Institute for Infocomm Research.*, 2002.