Transfer learning approach for financial applications

Cosmin Stamate
Dept. of Computer science
and Information Systems
Birkbeck, University of London
London, United Kingdom
c.stamate@dcs.bbk.ac.uk

George D. Magoulas
Dept. of Computer science
and Information Systems
Birkbeck, University of London
London, United Kingdom
gmagoulas@dcs.bbk.ac.uk

Michael S.C. Thomas
Dept. of Psychological Sciences
Birkbeck, University of London
London, United Kingdom
m.thomas@psychology.bbk.ac.uk

Abstract—**Artificial neural networks learn how to solve new problems through a computationally intense and time consuming process. One way to reduce the amount of time required is to inject pre-existing knowledge into the network. To make use of past knowledge, we can take advantage of techniques that *transfer the knowledge* learned from one task, and reuse it on another (sometimes unrelated) task. In this paper we propose a novel selective breeding technique that extends the transfer learning with behavioural genetics approach proposed by Kohli, Magoulas and Thomas (2013), and evaluate its performance on financial data. Numerical evidence demonstrates the credibility of the new approach. We provide insights on the operation of transfer learning and highlight the benefits of using behavioural principles and selective breeding when tackling a set of diverse financial applications problems.**

Keywords—*transfer learning, artificial neural networks, genetic algorithms, population studies, behavioural genetics, selective breeding, financial applications*

I. INTRODUCTION

It is fundamental for financial institutions to provide accurate risk assessments when evaluating new and/or existing customers (applicants). This risk assessment process should be continuous, evolve with the customer and/or institution requirements [1]. Most of the traditional assessment techniques that financial institutions use include but are not limited to: artificial neural networks (ANNs), support vector machines and complex tree structure predictors [2]. It is wasteful to use these techniques in isolation as it does not exploit inter-departmental and/or inter-institutional sharing of *valuable* information (knowledge). What we are proposing in this paper is a novel framework that has its roots in behavioural genetic studies, and provides a technique that facilitates the use of existing customer knowledge (from different departments or even between institutions) to improve the overall efficiency and speed of traditional customer evaluation techniques. This framework stems from previous work done by Kohli et al. [3], [4], with the difference that it uses artificial breeding techniques between its populations of genetic algorithms to make this transfer of learning efficient on data with a financial aspect.

Genetic algorithms are artificial intelligence's search heuristic which can easily mirror the process of natural selection. Sub-branch of the vast evolutionary algorithms class that are inspired from the natural processes of evolution [5]. For genetic algorithms to behave in an evolutionary way we need to implement, all of the following evolutionary processes: selection, mutation, inheritance and crossover [6]. In order to take full advantage of the power of evolutionary algorithms we need to use populations of abstractions (ANNs as individuals) which over iterations (time) will evolve towards better solutions; this will happen without the need of any external interference. Traditionally we start from random populations of individuals and iteratively improve on the quality of the population by evaluating each individuals fitness, selecting only the best performing and mating them. We stop the evolution when we reach a generation threshold or when we hit the fitness target for the population [5].

One of the best performing classifiers of financial data used by financial institutions have been ANNs [1]. These have been tried and tested against statistically sound methods, and provide one of the most robust approaches to perform large-scale classification of financial data known to date [2]. We have managed to use genetic algorithms (with novel selection and mating techniques) and traditional feed-forward ANNs and form a hybrid algorithm. This hybrid *behavioural genetics* inspired algorithm is successful in performing transfer learning on heterogeneous tasks, and is also used to asses, to some degree, task similarity [4].

In transfer learning the goal is to transfer useful knowledge from a source task (input) to a target task (output). We differentiate between two types of transfer paradigms: functional, where learning happens simultaneously in the source and in the target; and representational, where tasks are learned at different times [7], [8]. Most transfer learning research assumes a relationship between source and target tasks [9], [7], [8]. Therefore the problem is that most of the transfer learning algorithms expect source and target tasks to be similar, otherwise they will not produce positive results [4], [7], [8]. Transfer learning between different tasks poses high risk as it might result in negative transfer, which damages learning capabilities. In traditional approaches, the risk of transfer having a negative impact on learning and performance is dictated by the degree of similarity between tasks [4], [7], [8]. We focus on the branch of transfer learning which deals with harnessing the ability to learn. We move the top performing lerners from generation to generation, until our learning models are very optimised for learning. By combining specific breeding techniques with the population based behavioural genetics framework we have achieved positive transfer learning on diverse financial tasks.

The paper is structured as follows. In the next section we give an overview of *behavioural genetics* and explain the main ideas that underpin our approach. In Section III the proposed methodology for transfer learning is derived and

in Section IV the financial datasets are described. Section V presents experimental results. The paper ends with concluding remarks in Section VI.

II. Overview of behavioural genetics

Behavioural genetics studies the inheritance of traits between individuals of the same population, and then translates them into similarities and/or differences. Individual variation is separated in genetic and environmental components, which are examined for their influences on animal (including human) behaviour. In animal studies breeding and gene knockout techniques are commonly used to extract the required information. For humans the use of twin or adoption studies is more common. In this paper we have combined both human and animal studies by using both selective breeding and twin studies.

We have to stress the importance of the environment in individual development, this is simulated with the help of a filter applied to each training set. To get a better appreciation of the effect genetic and environmental influences have on network performance, we use the research on english past tense acquisition, with populations of identical and fraternal twins [3]. To provide better variation within populations, we use monozygotic (MZ, genetically identical twins, 100% same genetic material) and dizygotic (DZ, genetically similar or fraternal twins, 50% same genetic material), which in ANN terms translate to networks with identical hyper-parameters for MZ and 50% identical hyper-parameters for DZ. Both type of networks (MZ or DZ) have random initial weights between their hidden nodes [4].

Environmental influences are simulated by a unique filter which is applied to the training set for each individual of the population. The filtered training set resembles the notion of socio-economic-status (SES) [4]. Research on the effect of SES on language acquisition shows that children with lower SES will perform significantly worse at the same task than children with a higher SES. This happens because of differences in the amount and quality of learning resources [10]. We ensured that the equal environment assumption was kept in all the simulated populations, which means that both MZ and DZ individual twins share the same environment [3], [4]. Splitting the populations between 50% MZ and 50% DZ twins, not only ensures good variability but proves helpful when attempting to learn multiple unrelated tasks [3].

By incorporating and optimizing selective mating techniques into the behavioural genetics framework outlined above we have managed to focus and enhance the transfer of learning obtained on heterogeneous tasks [4]. We got the inspiration from the animal kingdom where certain animals have been mated (and still are now) for specific traits by humans, throughout history [11].

III. Proposed Transfer Learning Methodology

In this section we present the key components of our approach. This is based on a modification of the behavioural genetics population framework proposed by[4] that is combined with a new selective breading technique to enhance transfer of learning. Before we can use these algorithms we need to

homogenise the feature dimensionality of our datasets, transforming them into environments with the same dimensionality for learning algorithms but keeping the internal distribution of the underlying problem the same. This is a requirement because we want our populations and individuals to be able to change environments without any external intervention on population and/or individual topology. Next we divide the datasets into training, validation and testing, using a 60 - 20 - 20 distribution and we also split the data equally on the classes we have. This is useful not only from a machine learning context but it also mimics training with a curriculum [12]. It also ensures that training, validation and testing will have the same distribution of class examples and still have elements of randomness.

Algorithm 1 takes inspiration from the behavioural genetics framework developed by Kohli [4] but does not place any emphasis on heritability. A big part of Kohli's behavioural genetics framework is dedicated to heritability, which is abstracted, to some degree to task relatedness. We have used a more traditional approach to determine task relatedness, as outlined in the next section. SES has been randomly applied to each individual, from every population in the training phase, by removing vectors from the training set. SES is random and set between 0 and 40%, the latter meaning that 40% of the training set will be removed.

A novel selective breeding technique is presented in Algorithm 2. The selection part of this algorithm has no natural basis and has been developed in this way to fully exploit the variability and relationship between best and medium level performing twins. The best performing members of the population are selected for accuracy and medium level are selected to encourage population flexibility. Mixing mid level individuals with top performers is important for positive transfer of learning, otherwise, by choosing only top performers we have populations that are focused on optimal results and not on transfer. By mating top performers with medium level we obtain a good synergy between optimum results and transfer flexibility. The crossover and mutation part of the algorithm is the standard genetic algorithm random single point crossover and 0.1% chance of random mutation.

IV. Datasets

All the sets of data used in this work have been selected from the UC Irvine Machine Learning Repository [13] and are comparable with data published in [14], [15], [1], [16]. Two of the three datasets are industry standards when it comes to assessing classifier accuracy on financial data [14], [1], [16], and have been used to assess transfer of learning in [15]. A short description of the datasets is available below and a comparison between characteristics in is shown in Table I.

- **Statlog (Australian Credit Approval)**[1] (Australian) – credit card applications with a good mix of attributes: continuous, small and large numbers of nominal values plus some missing values. This dataset was used in [17], [18], [15], [14].

- **Statlog (German Credit Data)**[2] (German) – credit data provided by Prof. Hofmann and the numerical

---

[1]https://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval)
[2]https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

Algorithm 1 Hybrid algorithm that exploits behavioural genetics intuitions in a transfer learning context

Require: $datasets \leftarrow$ load all the preprocessed data-sets here
1: **for** $dataset$ in $datasets$ **do**
2:    $populations \leftarrow$ initialize 2 populations of the same size with random values (between specific intervals, pre-determined by the distribution of the data-set) for the parameters of each individual
3:    **while** $populations$ threshold is not reached **do** ▷ we have used 20 generations as our threshold
4:       $brothers \leftarrow emptyArray$
5:       **for** $population$ in $populations$ **do**
6:          train($population$, $dataset$) ▷ only on training and validation sets, this also applies a random SES to the training set
7:          $fitness \leftarrow$ assess($population$, $dataset$) ▷ on testing set only
8:          $leaders \leftarrow$ selectForMating($population$, $fitness$) ▷ extract top and middle performers from the population
9:          $newPopulation \leftarrow$ mate($leaders$) ▷ custom breeding method, outlined at algorithm 2
10:         $brothers \leftarrow$ split($newPopulation$) ▷ split them into 2 equal populations (arrays of individuals) so that there are no identical twins in the same population
11:       **end for**
12:       $populations \leftarrow$ combine($brothers$) ▷ create 2 populations (arrays) from the 4 currently residing in $brothers$ by merging the populations that will not share any identical twins between them
13:    **end while**
14: **end for**

Algorithm 2 Selective breeding algorithm for transfer of learning

Require: $top \leftarrow$ top performers from a population
Require: $mid \leftarrow$ average performers from a population
1: $children \leftarrow$ crossover(top) ▷ combine all the top performers in pairs and produce 4 offspring for each pair: 2 MZ twins, 2 DZ twins
2: $children \leftarrow$ crossover(top, mid) ▷ combine middle and top performers in pairs and produce 4 offspring for each pair: 2 MZ twins, 2 DZ twins
3: $children \leftarrow$ crossover(mid) ▷ combine all the middle performers in pairs and produce 4 offspring for each pair: 2 MZ twins, 2 DZ twins

---

dataset provided by Strathclyde University. Used in [14], [15], [1], [16]

- **Banknote Authentication**[3] (Banknote) – data extracted from real images of forged banknotes, with the help of an industrial camera. Provided by Volker Lohweg (University of Applied Sciences, Ostwestfalen-Lippe). Used in [19].

As you can see two of the datasets are from the SatLog project. The SatLog project involves comparing the performances of machine learning, statistical, and ANNs algo-

rithms on data sets from real-world industrial areas including medicine, finance, image analysis, and engineering design. The two datasets chosen by us from the SatLog project are popular in the machine learning and the financial world.

Datasets	Inputs	Instances	Attribute types
Australian	14	690	6 numerical and 8 categorical
German	24	1000	numerical
Banknote	5	1372	numerical

TABLE I.   DIMENSIONALITY OF FEATURE SPACE WITH ATTRIBUTE TYPES. ALL TARGETS HAVE 2 CLASSES.

The networks had the same number of hidden nodes for each task (the other intrinsic parameters were optimally selected), and the mean weight difference calculated between the various tasks, as can be seen in Table II.

To assess task relatedness on the training, validation and testing sets of data for each task, we have used the mean difference between the weight space of best performing ANNs of the same size (same number of input, hidden and output nodes and layers) [7]. We have chosen a fixed 100 hidden nodes ANN, and with the rest of the hyper-parameters (learning rate, momentum and slope of logistic function) varying to perform optimally on each of the different tasks. After finding the best hyper-parameters for each of our 3 datasets, we get the mean of the norm of the difference of the weight vectors between our datasets. You can see the outcome in Table II: the smaller the difference the more related the tasks are.

Datasets	Australian	German	Banknote
Australian	0.0000	0.0426	0.0448
German	0.0426	0.0000	0.0401
Banknote	0.0448	0.0401	0.0000

TABLE II.   THE MEAN OF THE NORM OF THE DIFFERENCE OF THE WEIGHT VECTORS.

V. EXPERIMENTS AND RESULTS

Before any of the experiments can commence we have to make sure that the data is in the format described in section III. We need the data normalized and split into training, validation and testing sets.

We have two main sections for results, firstly we got the optimal neuro-computational calibrations for each dataset. This helped us encode each source task with its optimal ranges and it has enabled us to narrow down the initial search space. Although we can get the best set of hyper-parameters, we are not after best performance, what we are interested in is using intervals that center around the best performing hyper-parameters. This will ensure that transfer of learning can occur as the source networks will not be *focused* only on the source task, and will allow a smooth transition of *learning* from source to target(s). Sometimes the best hyper-parameter cannot be interpreted as the center of an interval, in that case we take the lowest most significant neuro-computational parameter as lower bound of the interval and the highest as the other. We have chosen only 4 hyper-parameters to optimise: number of hidden nodes, learning rate, momentum and slope of logistic function, but many more can be incorporated in this framework. You can see the calibration bounds of the datasets in Table III.

Datasets	Hidden nodes	Learning rate	Momentum	Logistic slope
Australian	15 to 50	0.01 to 0.2	0.01 to 5.1	1 to 4
German	5 to 30	0.01 to 0.4	0.1 to 1.2	0.8 to 2.1
Banknote	5 to 15	0.01 to 0.15	0.01 to 0.01	0.01 to 1.2

TABLE III.   OPTIMAL SOURCE TASK CALIBRATIONS.

After we have the source calibrations, we start using the two algorithms described in section III. For the purposes of these experiments we have used populations of 1200 individuals (ANNs), each trained for 1000 epochs, without early stopping [20]. The fitness criterion was overall miss-classification error on the test data set. We have started with two random (but within the bounds presented in table III) populations. We select a source task, from the 3 available and train the 2 random populations on this chosen task. Furthermore we evolve these populations by following the steps outlined in Algorithm 1 and Algorithm 2 for 20 generations. This produced highly optimised populations for the chosen source task, and populations that are flexible enough to transfer the acquired learning knowledge to aid learning of new tasks. The flexibility is possible because of the selective breeding pressures outlined in Algorithm 2. We take the 2 populations, breed them into one using the same technique as before (get top and middle performers and breed them) but without producing MZ and DZ twins. This produces only 2 non-identical offspring per pair and get a 1200 individual population, which is the population that best encodes the selected source task. We do the same for the remaining 2 source tasks, and have now one 1200 individuals population optimised for each of our tasks (datasets).

The optimised populations are then trained and tested on each of the tasks, and the results are presented in Tables IV, V and VI. The top part of each table represents the mean performance of the population on the validation data set of each target task, and the lower represents performance on the test set for each target task. When compared with results from literature [14], [1], [16], this approach is not the best performing one. This is because we have focused our efforts in producing positive transfer of learning, without any negative impact on the overall process of learning. Whilst approaches in [1], [14], [16] were conceived to solve one of the problems outlined in the datasets in isolation. We have succeeded in solving all problems together and share knowledge between solutions, as one can see from the continuity of the results in Tables IV, V and VI. The continuity is not only in the overall classification error from source to target tasks but also in all the underlying parts that make that error. As one can see there is no spike or abrupt jump in values between all the true and false positives or negatives. When comparing our approach with a population (1200 individuals) of randomly initialised individuals (networks), we can see that our approach is considerably more accurate. This comparison becomes evident in Figure 1 with reinforcing testing and validation results presented in Table VII.

Targets	Australian	German	Banknote
True positives	330.75	137.5	737.25
True negatives	267.5	606.5	607.75
False positives	39.5	93.5	2.25
False negatives	52.25	162.5	24.75
Precision	0.837	0.79	0.961
Recall	0.871	0.866	0.996
Mean square error	13.297	25.6	1.968
True positives	329.75	138.0	736.25
True negatives	264.5	608.75	607.25
False positives	42.5	91.25	2.75
False negatives	53.25	162.0	25.75
Precision	0.833	0.790	0.959
Recall	0.862	0.870	0.995
Mean square error	13.877	25.325	2.077

TABLE IV.   EVALUATIONS ON THE VALIDATION (UPPER) AND TESTING (LOWER) DATASETS FOR SOURCE TASK: **AUSTRALIAN**.

Targets	Australian	German	Banknote
True positives	327.5	138.75	737.75
True negatives	268.75	624.0	608.25
False positives	38.25	76.0	1.75
False negatives	55.5	161.25	24.25
Precision	0.829	0.795	0.962
Recall	0.875	0.891	0.997
Mean square error	13.587	23.725	1.895
True positives	330.75	127.0	737.5
True negatives	270.0	619.5	608.75
False positives	37.0	80.5	1.25
False negatives	52.25	173.0	24.5
Precision	0.838	0.782	0.961
Recall	0.879	0.885	0.998
Mean square error	12.935	25.35	1.877

TABLE V.   EVALUATIONS ON THE VALIDATION (UPPER) AND TESTING (LOWER) DATASETS FOR SOURCE TASK: **GERMAN**.

Targets	Australian	German	Banknote
True positives	337.5	111.75	738.75
True negatives	263.75	630.5	607.25
False positives	43.25	69.5	2.75
False negatives	45.5	188.25	23.25
Precision	0.853	0.773	0.963
Recall	0.859	0.901	0.995
Mean square error	12.862	25.775	1.895
True positives	334.25	136.75	735.5
True negatives	262.25	621.0	607.25
False positives	44.75	79.0	2.75
False negatives	48.75	163.25	26.5
Precision	0.843	0.793	0.958
Recall	0.854	0.887	0.995
Mean square error	13.550	24.225	2.132

TABLE VI.   EVALUATIONS ON THE VALIDATION (UPPER) AND TESTING (LOWER) DATASETS FOR SOURCE TASK: **BANKNOTE**.

Targets	Australian	German	Banknote
Validation error	28.155	33.3	2.22
Testing error	37.5	28.7	3.01

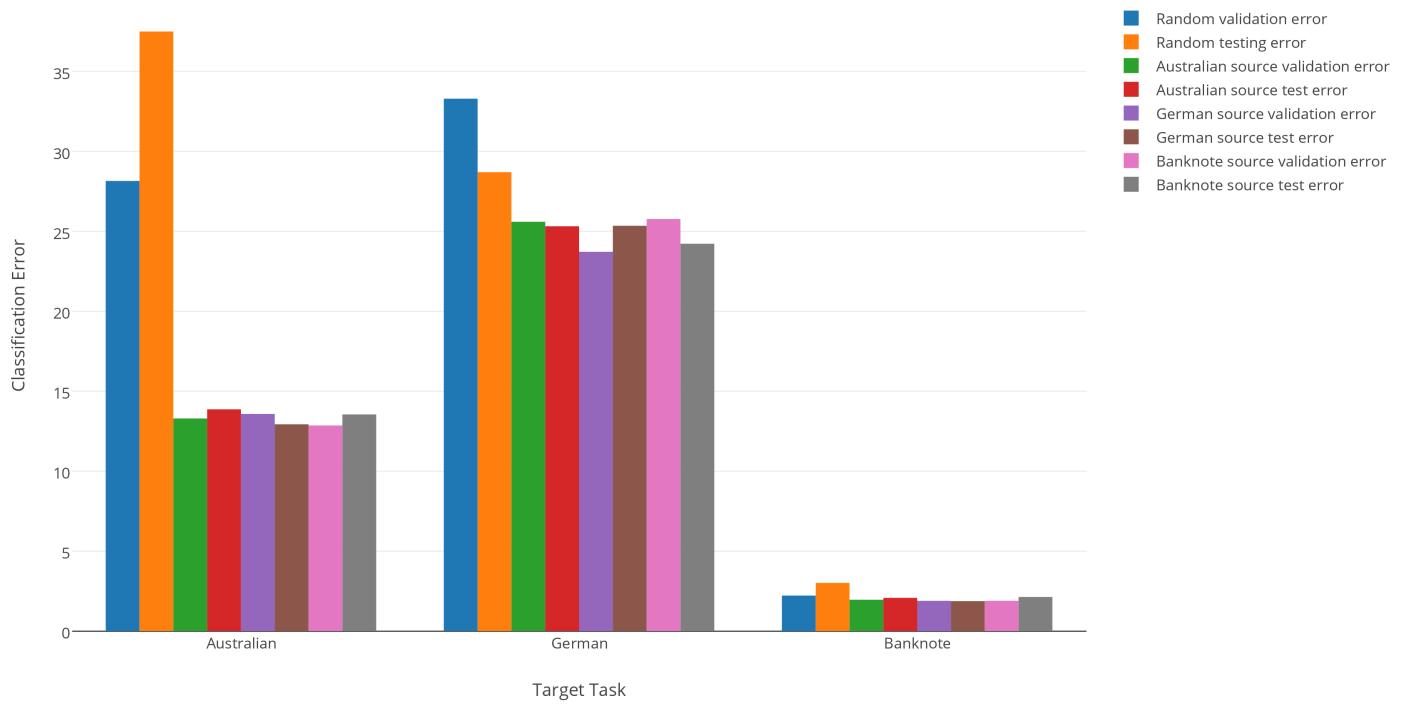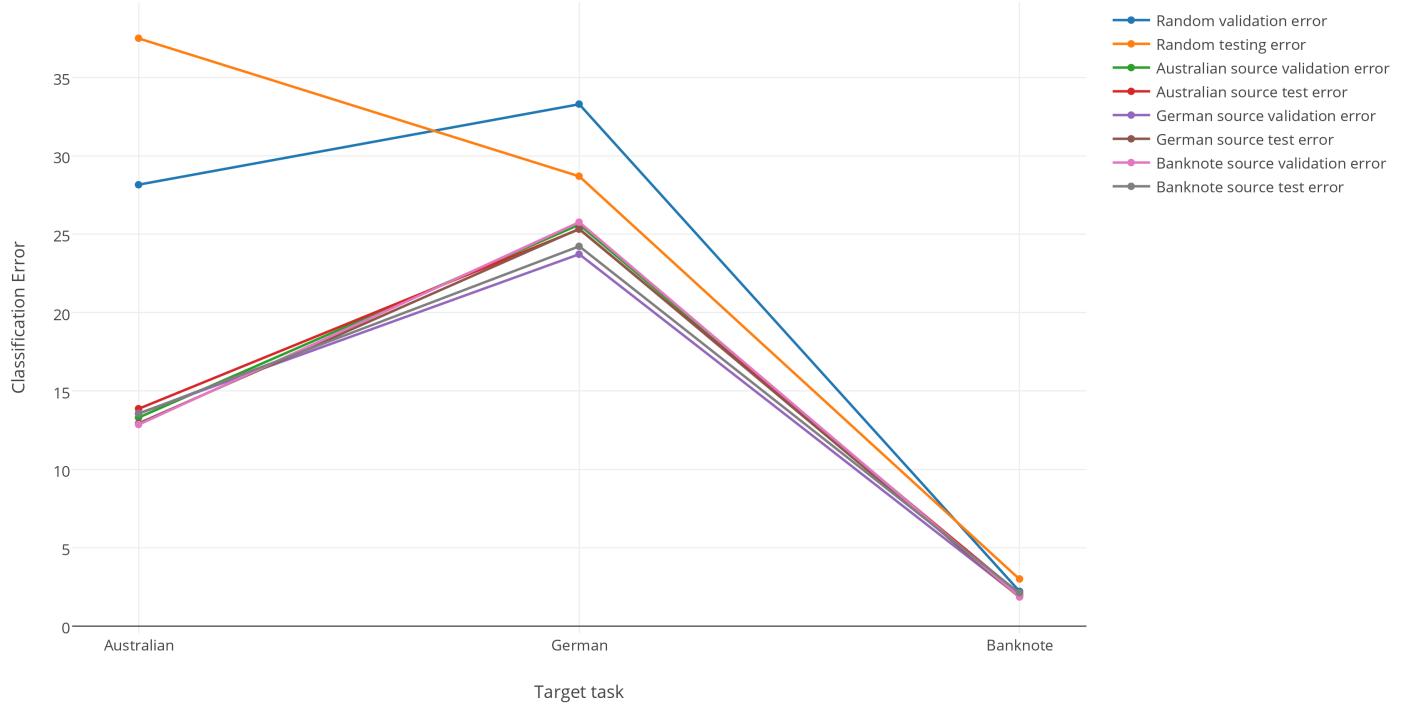TABLE VII.   CLASSIFICATION ERROR FOR RANDOMLY INITIALISED NETWORKS

Fig. 1. Benchmark between randomly initialised networks and our proposed transfer learning approach

VI. Conclusion

Diverse financial data has been chosen as the topic of this work because financial institutions have and still are producing large amount of data on a daily basis. In contrast to approaches proposed in the literature where special methods are developed to solve a particular problem, we propose a transfer learning approach that exploits knowledge acquired in previous situations to learn a new problem. Approaches based on transfer learning are normally affected by negative transfer. The proposed framework based on behaviour genetics offers flexibility when learning different problems, alleviating the issue of negative transfer in the proposed datasets. Furthermore, by using selective breeding, like humanity has done with animals for thousands of years, we have managed to improve overall accuracy and keep the learning flexibility. Novel selective breeding techniques injected in an already successful behavioural genetics computational framework, resulted in optimised positive transfer of learning. Although the results that we have reported are good from a classification and transfer point of view we could still improve this approach by utilizing more sophisticated selective breeding techniques and ensemble generation from populations based on individual class accuracy. In future work we intend to explore these methods in more depth.

References

[1] A. Khashman, "Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes," *Expert Systems with Applications*, vol. 37, no. 9, pp. 6233–6239, sep 2010. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2010.02.101

[2] J. N. Crook, D. B. Edelman, and L. C. Thomas, "Recent developments in consumer credit risk assessment," *European Journal of Operational Research*, vol. 183, no. 3, pp. 1447–1465, dec 2007. [Online]. Available: http://dx.doi.org/10.1016/j.ejor.2006.09.100

[3] M. Kohli, G. D. Magoulas, and M. S. C. Thomas, "Hybrid computational model for producing english past tense verbs," in *EANN*, ser. Communications in Computer and Information Science, C. Jayne, S. Yue, and L. S. Iliadis, Eds., vol. 311. Springer, 2012, pp. 315–324.

[4] ——, "Transfer learning across heterogeneous tasks using behavioural genetic principles," in *UKCI*. IEEE, 2013, pp. 151–158.

[5] L. Davis, Ed., *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, 1991.

[6] T. Back, *Evolutionary algorithms in theory and practice - evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press, 1996.

[7] L. Y. Pratt, "A survey of transfer between connectionist networks," *Connect. Sci.*, vol. 8, no. 2, pp. 163–184, 1996.

[8] S. Thrun and L. Pratt, *Learning to Learn*. Kluwer Academic Publishers, 1998.

[9] A. Robins, "Transfer in cognition," *Connect. Sci.*, vol. 8, no. 2, pp. 185–204, 1996.

[10] M. S. C. Thomas, N. A. Forrester, and A. Ronald, "Modeling socioeconomic status effects on language development." *Developmental Psychology*, vol. 49, no. 12, pp. 2325–2343, 2013. [Online]. Available: http://dx.doi.org/10.1037/a0032301

[11] Y. Jiang, D. I. Bolnick, and M. Kirkpatrick, "Assortative mating in animals," *The American Naturalist*, vol. 181, no. 6, pp. E125–E138, jun 2013. [Online]. Available: http://dx.doi.org/10.1086/670160

[12] Y. Bengio, "Learning deep architectures for ai." *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009. [Online]. Available: http://dblp.uni-trier.de/db/journals/ftml/ftml2.htmlBengio09

[13] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[14] C. TSAI and J. WU, "Using neural network ensembles for bankruptcy prediction and credit scoring," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2639–2649, may 2008. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2007.05.019

[15] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[16] Z. Zhao, S. Xu, B. H. Kang, M. M. J. Kabir, Y. Liu, and R. Wasinger, "Investigation and improvement of multi-layer perceptron neural networks for credit scoring," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3508–3516, may 2015. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2014.12.006

[17] J. R. Quinlan, "C4.5: Programs for machine learning," in *Morgan Kaufmann series in machine learning.*, 1993.

[18] ——, "Simplifying decision trees," *Int. J. Hum.-Comput. Stud*, vol. 51, p. 497, 1999.

[19] V. Lohweg, J. L. Hoffmann, H. Dörksen, R. Hildebrand, E. Gillich, J. Hofmann, and J. Schaede, "Banknote authentication with mobile devices," in *Media Watermarking, Security, and Forensics 2013*, A. M. Alattar, N. D. Memon, and C. D. Heitzenrater, Eds. SPIE, feb 2013. [Online]. Available: http://dx.doi.org/10.1117/12.2001444

[20] N. Morgan and H. Bourlard, "Generalization and parameter estimation in feedforward nets: Some experiments," International Computer Science Institute, Berkeley, CA, Tech. Rep. TR-89-017, 1989.