An expert system for extracting knowledge from customers' reviews: The case of Amazon.com, Inc.

Mauro Castelli[a], Luca Manzoni[b], Leonardo Vanneschi[a], Aleš Popovič[a,c,*]

[a] NOVA IMS, Universidade Nova de Lisboa, 1070-312, Lisboa, Portugal
[b] DISCo, Dipartimento di Informatica Sistemistica e Comunicazione, University of Milano Bicocca, 20126, Milano, Italy
[c] Faculty of Economics, University of Ljubljana, Kardeljeva Ploščad 17, 1000 Ljubljana, Slovenia

ARTICLE INFO

Accepted 5 May 2017

Keywords:
Genetic programming
Semantics
E-commerce
Customers' feedback

ABSTRACT

E-commerce has proliferated in the daily activities of end-consumers and firms alike. For firms, consumer satisfaction is an important indicator of e-commerce success. Today, consumers' reviews and feedback are increasingly shaping consumer intentions regarding new purchases and repeated purchases, while helping to attract new customers. In our work, we use an expert system to predict the sentiment of a product considering a subset of available customers' reviews.

1. Introduction

In the marketing and management literature, customer satisfaction is increasingly emphasized as a vital factor for increasing sales and thus corporate performance (Anderson, Fornell, & Lehmann, 1994; Balasubramanian, Konana, & Menon, 2003; Szymanski & Henard, 2001). As firms engage ever more in e-commerce initiatives, customer satisfaction is an important indicator of e-commerce success. Despite the growth in sales and volumes seen around the world, e-commerce sites often tend to underestimate customer reviews' importance for new purchases and repeat purchases, as well as for attracting new customers. More often than not, customer reviews are neglected on the list of what firms believe are the critical success factors of an e-commerce initiative (Liu & Arnett, 2000). In fact, firms tend to focus chiefly on optimizing the design of the website, customer service and support, and the administrative activities related to the management of e-commerce (Bendoly & Kaefer, 2004; Bergendahl, 2005; Teo & Liu, 2007).

Although all of these activities require valuable effort, overlooking customer reviews and feedback can curtail the success of e-commerce (Cui, Lui, & Guo, 2012; Kim, Galliers, Shin, Ryoo, & Kim, 2012; Lee, Han, & Suh, 2014; Qiu, Lin, & Li, 2015). More specifically, according to the "Social Commerce2007" report of Bazaarvoice (Bazaarvoice, 2007) as a result of the opportunity given to customers to provide feedback and reviews on purchased products, 42% of e-commerce managers reported a significant rise in the volume and average order value. On the other hand, only 6% of e-commerce managers revealed a decline in orders after the reviews were made public. Similarly, large e-commerce sites such as Amazon and eBay have found that many consumers appreciate the opportunity to evaluate products before their (possible) purchase. About 40% of customers participating in a survey by Nielsen (Nielsen, 2010) stated they would not have purchased an electronic item without having had access to the opinions of other customers, whereas 85.57% of the participants said they read reviews often or very often before buying online. In this perspective, an additional factor with a significant impact on online purchasing behavior is the customer social network. As reported in Verbraken, Goethals, Verbeke, and Baesens (2014), knowledge of a person's social network can help in predicting that person's e-commerce acceptance of different products. The quality of reviews is also considered to be very important. A study reported in Lackermair, Kailer, and Kanmaz (2013) found that 75% of customers reported that the quality of such reviews greatly influences their decision to purchase a product from an online store. Last but not least, the possibility to use comments made by customers in order to improve the SEO (Search Engine Optimization) process also needs to be considered as a relevant factor for firms. In fact, the content of customers' opinions could be indexed and used to produce search engine results. In this case, the advantage is that the reviews are written in

* Corresponding author.
E-mail addresses: mcastelli@novaims.unl.pt, castelli.mauro@gmail.com (M. Castelli), luca.manzoni@disco.unimib.it (L. Manzoni), lvanneschi@novaims.unl.pt (L. Vanneschi), apopovic@novaims.unl.pt (A. Popovič).

natural language, which is able to match many keywords and thus give a further boost to the SEO task.

Obviously, customer feedback cannot be considered the only variable that can (positively or negatively) affect the average value of orders. Thus, firms need to consider the broader interplay of factors to fully comprehend which enable and which inhibit e-commerce success (Gefen, 2000). An analysis of the role of all possible components is very complex, yet, the strong correlation between reviews and firm sales, as well as studies demonstrating the importance of reviews in establishing an e-commerce website's reliability, make e-commerce sites' inclusion of customer reviews an established practice.

The ability to predict the score of future reviews is useful in many applications: for instance, it is possible to suggest, among objects with similar ratings, the one that has the highest expected future review score. It may also be useful for predicting issues with the items: from the detection of sellers with counterfeit objects to issues concerning the manipulation of the reviews (Hu, Liu, & Sambamurthy, 2011). In both cases, a score that varies too much with respect to the predicted one can be interpreted as a signal of a possible anomaly. Other studies dealing with the importance of predicting the review score of an item are presented in Qu, Ifrim, and Weikum (2010), Gupta, Di Fabbrizio, and Haffner (2010) and Ganu, Elhadad, and Marian (2009).

All aspects mentioned so far show that customer feedback is an important asset for e-commerce managers. Hence, extracting non-trivial knowledge and manageable information from such rich data pools is a challenging issue of paramount importance for e-commerce managers.

To answer this call, in this paper we propose the use of a machine learning (ML) technique. The application of a ML technique tries to overcome the limitations of traditional statistic-based linear regression methods. Although these techniques and models are reliable, they are the best choice in managing unstructured data or data where no previous knowledge of the underlying model is available. Hence, more sophisticated means must be employed to extract meaningful information from data. ML methods have shown an ability to perform better when dealing with non-linearity and unstructured and complex data. While existing ML techniques have been successfully used to address problems in different domains, researchers continuously seek to advance existing methods and provide novel ones for analyzing data sets to make sense of the data, extract useful information, and build knowledge to inform decision-making. In this light, and considering the large amount of data available today, in this paper we propose an artificial intelligence system for extracting useful information considering the feedback of e-commerce customers. The proposed algorithm is a variant of the standard genetic programming (GP) algorithm but, unlike the standard one, it is able to scale beyond data sets of a few million elements and it is based on a solid theoretical background that guarantees the existence of certain properties that will help the search process produce more reliable solutions

The paper is organized as follows: Section 2 describes the standard GP algorithm and the operators used in the search process. Section 3 presents the geometric semantic operators used in this paper. More specifically, we highlight the benefits of the operators on the search process. Section 4 describes the experimental phase and discusses the results obtained. Section 5 concludes the paper, providing some directions for possible future research.

2. Genetic programming

Genetic Programming (GP) is a technique that comes from a larger computational intelligence research area called evolutionary computation (EC). GP consists of the automated learning of computer programs by means of a process inspired by biological evolution (Koza, 1992). Generation by generation, GP stochastically transforms populations of programs into new, hopefully improved, populations of programs. The quality of a solution is expressed by using an objective function. The value of this objective function is the fitness of an individual. The search process of GP is shown in Fig. 1.

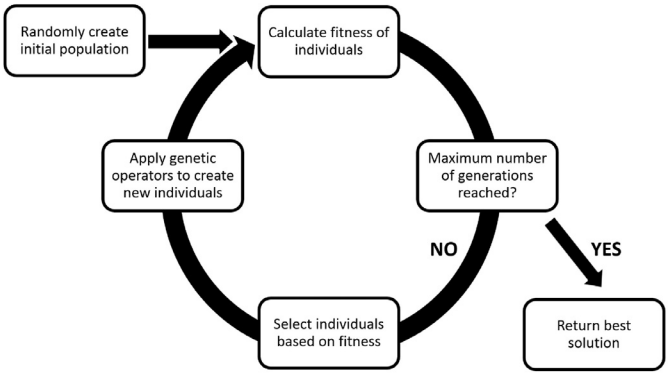

Fig. 1. The GP algorithm.

In order to transform a population into a new population of candidate solutions, GP uses particular operators called genetic operators. Considering the common tree representation of GP individuals, the standard genetic operators (crossover and mutation) act on the structure of the trees that represent the candidate solutions. In other terms, standard genetic operators act on the syntax of the programs. In this paper, we used genetic operators that, unlike the standard ones, are able to act at the semantic level. The definition of semantics used in this work is that which was also proposed in Moraglio, Krawiec, and Johnson (2012) and will become clear in the following section.

To understand the differences between the genetic operators used in this work and those used in the standard GP algorithm, the latter are briefly described. The "standard" crossover operator is traditionally used to combine the genetic material of two parents by swapping part of one parent with part of the other. In more detail, after choosing two individuals based on their fitness, the crossover operator performs two operations: 1) it selects a random subtree in each parent; and 2) swaps the selected subtrees between the two parents (the resulting individuals are the children). The mutation operator introduces random changes in the structures of individuals in the population. The best known mutation operator, called sub-tree mutation, works as follows: 1) it randomly selects a point in a tree; 2) it removes whatever is currently at the selected point and whatever is below that point; and 3) it inserts a randomly generated subtree at that point. This operation is controlled by a parameter that specifies the maximum size (usually measured in terms of tree depth) for the newly created subtree that is to be inserted.

3. Geometric semantic operators

Despite the large number of human-competitive results achieved with the use of GP (Koza, 2010), researchers continue to investigate new methods in order to improve GP's ability to produce optimal or quasi-optimal solutions. In recent years, an emerging idea is to include the concept of semantics in the evolutionary process performed by GP. While several studies exist (i.e. Beadle & Johnson, 2009; Castelli, Vanneschi, & Silva, 2014; Vanneschi, Castelli, & Silva, 2014a), the definition of semantics is not unique and this concept is interpreted in different ways and according to different perspectives. In this work, we use the most common and widely accepted definition of semantics. Hence, the

term semantics is used to refer to the behavior of a program once it is applied to a set of data. This definition relates the term semantics with the vector of outputs obtained after applying a given program (or candidate solution) to a set of training data (Moraglio et al., 2012). Including the semantics concept in the search process allows GP to overcome one of its current limitations. In fact, while semantics determines what a program actually does, the traditional genetic operators manipulate programs only in terms of their syntax. Hence, traditional GP operators completely ignore the information about the behavior of programs provided by semantics. The drawback of this choice is that it is difficult (or even impossible) to predict the effect modifications which affect the syntax of the programs will have on the semantics of the same programs.

To overcome this problem, new genetic operators that act on the semantics of the programs have recently been defined (Moraglio et al., 2012). In particular, among the several advantages of these operators with respect to the traditional ones, it has been shown that these operators are able to induce a unimodal fitness landscape (Stadler, 1995) on any problem entailing finding a match between a set of input data and a set of expected target ones. In this paper, we will consider the definition of geometric semantic operators for real functions' domains since these are the operators we will use in the experimental phase.

Geometric Semantic Crossover. Given two parent functions $T_1, T_2 : \mathbb{R}^n \to \mathbb{R}$, the geometric semantic crossover returns the real function $T_{XO} = (T_1 \cdot T_R) + ((1 - T_R) \cdot T_2)$, where T_R is a random real function whose output values range in the interval [0, 1].

The interested reader is referred to Moraglio et al. (2012) for formal proof of the fact that this operator corresponds to a geometric crossover on the semantic space in the sense it produces an offspring that stands between its parents in this space. Nevertheless, even without formal proof, we can have an intuition of it by considering that the (unique) offspring generated by this crossover has a semantic vector that is a linear combination of the semantics of the parents with random coefficients included in [0, 1].

To constrain T_R in producing values in [0, 1] we use the sigmoid function: $T_R = \frac{1}{1+e^{-T_{rand}}}$ where T_{rand} is a random tree with no constraints on the output values.

Geometric Semantic Mutation. Given a parent function $T : \mathbb{R}^n \to \mathbb{R}$, the geometric semantic mutation with mutation step ms returns the real function $T_M = T + ms \cdot (T_{R1} - T_{R2})$, where T_{R1} and T_{R2} are random real functions with codomain in the range [0, 1].

These operators are able to transform each regression problem in a problem characterized by a unimodal fitness landscape (Moraglio et al., 2012). This property guarantees the algorithms convergence to- wards optimal solutions and allows semantic GP to outperform the standard syntax-based GP on regression problems. An introduction to geometric semantic operators can be found in Vanneschi (2017).

While these operators have several advantages (as reported in Moraglio et al., 2012), there is an important limitation that must be considered. As can easily be noticed considering their definition, every application of these operators produces an offspring that contains the complete structure of the parents, plus one or more random trees as its subtrees and some arithmetic operators: the size of each offspring is thus clearly much larger than the size of its parents. In order to counteract this exponential growth of the individuals (Moraglio et al., 2012) that makes difficult to use these operators to address real-life problems, in this paper we use the solution proposed in Vanneschi, Castelli, Manzoni, and Silva (2013). In greater detail, the work described in Vanneschi et al. (2013) proposed a very simple and effective implementation of the GP algorithm that allows GP to use the geometric semantic operators in a feasible way. This is the implementation used in this paper and documented in Castelli, Silva, and Vanneschi (2014). As widely discussed in Vanneschi et al. (2013), this implementation and the features of the geometric operators allow the proposed system to be used to analyze large datasets within a reasonable amount of time. In particular, the system scales linearly with respect to the number of instances in the dataset.

4. Experiments

This section describes the business problem that was considered, the available data, the experimental settings, and the obtained results.

4.1. Problem and data

We used the datasets provided by McAuley and Leskovec (2013), which include all reviews appearing on amazon.com from June 1995 to March 2013. Each entry in the dataset is composed, among other fields, by an object identifier, the review score, the time of the review, and the usefulness of the review as a fraction $\frac{a}{b}$, where a is the number of people who have found the review useful and b the number of people who have evaluated the review. For each review, we considered the following feature:

- **Score**. The review score.
- **Usefulness**. The number of people who have evaluated the review and the number of people who have found the review useful, represented as a pair of numbers and referred to as the *usefulness* of the review.
- **Time**. The time (in days) between this review and the first review of the object to which the review refers.

We focused on predicting the average review score of an object given a limited number of reviews. For example, where an object has been reviewed 50 times, we use the first 10 reviews to predict the average of the last 40 reviews. Each entry in the dataset consists of reviews of a specific object. In our study, we only considered objects that have been reviewed at least 30 times (for size 10 and 20) or 50 times (for size 30 and 40). Only objects in two categories of the amazon store were considered:

- **The kindle store** (KS). This dataset, for size 10 and 20, consists of 923 objects (divided into a training set of 627 objects and a test set of 296 objects, a 70%–30% split). For size 30 and 40, the dataset consists of 554 objects (388 in the training set and 166 in the test set).
- **Industrial and Scientific** (IS). The resulting dataset consists of 311 objects (218 in the training set and 93 in the test set) for size 10 and 20. The dataset for size 30 and 40 consists of 212 objects (148 in training set and 64 in the test set).

4.2. Experimental settings

The settings used in the experiments are described here. As fitness, we used the Root Mean Square Error (RMSE) between the output of the GP individual and the corresponding target (i.e. expected output). In order to validate our results, each run uses only 70% of the data in the learning phase (the *training set*) and the remaining part is used for validation purposes (the test set). Specifically, each run considered a different partition of training and test instances. The results for the test set are particularly interesting since they represent the behavior of the algorithm on unseen data and can thus quantify the predictive ability of the generated models. In all the experiments, the population size consists of 100 individuals and each run was left to evolve for 1000 generations. The crossover probability was equal to 0.9, while the probability of mutation was 0.5, with a random mutation step as suggested

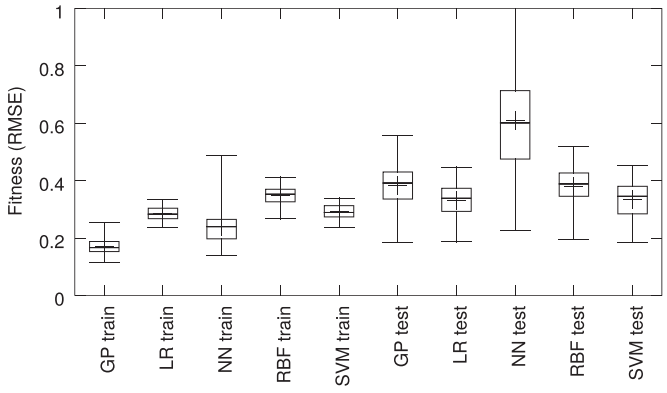

Fig. 2. The results on the IS dataset (S) for size 10.

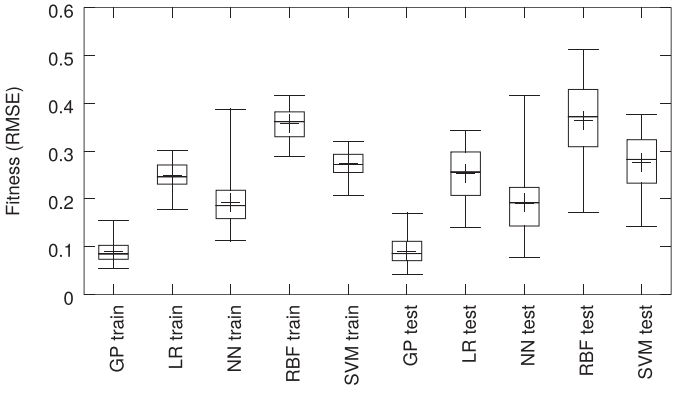

Fig. 3. The results on the IS dataset (S+U) for size 10.

in Vanneschi, Silva, Castelli, and Manzoni (2014b). The individuals are initialized using the ramped half-and-half method (Koza, 1992), with the maximal initial depth equal to 6. The functional operators were $+$, $\times$, $-$, and the protected division as in Koza (1992). The terminal nodes are the input variables and random constants in the range $[-100, 100]$. The values of these parameters were chosen after a preliminary tuning phase. In particular, several combinations of commonly used parameters' values were taken into account and, finally, we selected the set of parameters that returned the best performance. For both datasets considered, we performed 100 independent runs and we recorded, for each generation and for each run, the fitness of the best individual in the population in the training set, and the fitness of the same individual in the test sets. Each of the 100 runs was performed using a different split between the training and the test set. The results obtained were compared with those produced by other well-known state-of-the-art machine learning methods. This comparison allows us to draw some considerations about the competitiveness of the results. To perform the comparison between semantic GP with GSOs (hereinafter GSGP) and other machine learning methods, we used the implementations provided by the Weka public domain software (Weka Machine Learning Project, 2013). As done for GSGP, a preliminary study was performed in order to tune the considered techniques' parameters.

4.3. Results

Plots shown in Figs. 2–17 report the results achieved on the IS dataset. Denoting the interquartile range with *IQR*, the ends of the whiskers represent the lowest datum and the highest datum. The central bar denotes the median RMSE on the 100 runs performed and the cross represents the average. Several configurations were

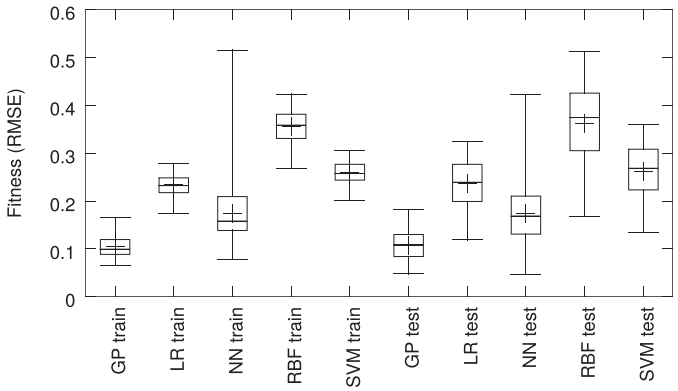

Fig. 4. The results on the IS dataset (S+U+T) for size 10.

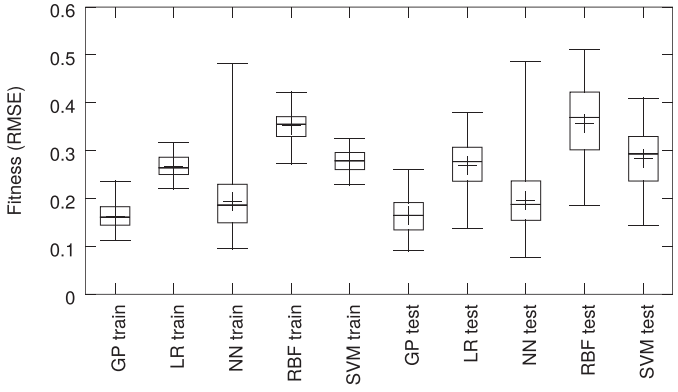

Fig. 5. The results on the IS dataset (S+T) for size 10.

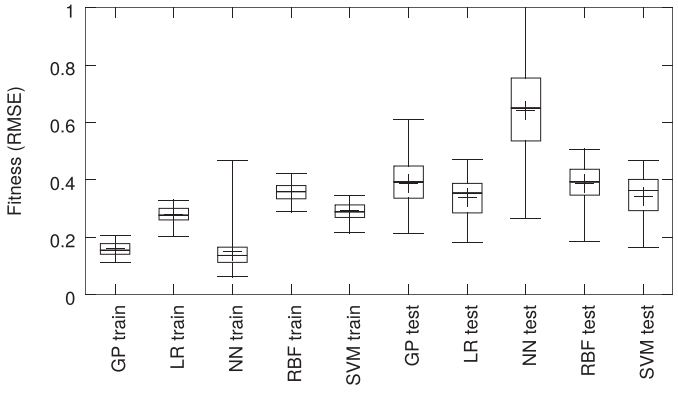

Fig. 6. The results on the IS dataset (S) for size 20.

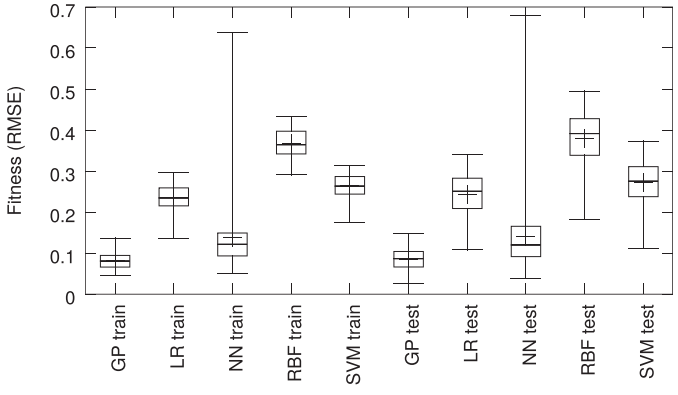

Fig. 7. The results on the IS dataset (S+U) for size 20.

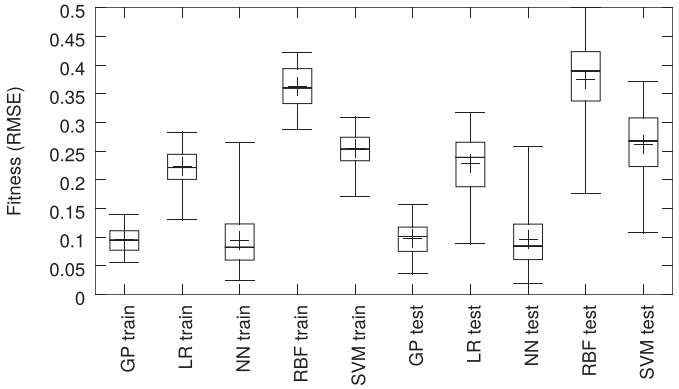

Fig. 8. The results on the IS dataset (S+U+T) for size 20.

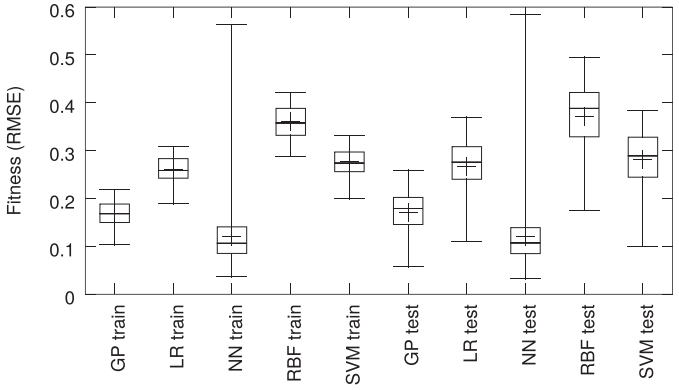

Fig. 9. The results on the IS dataset (S+T) for size 20.

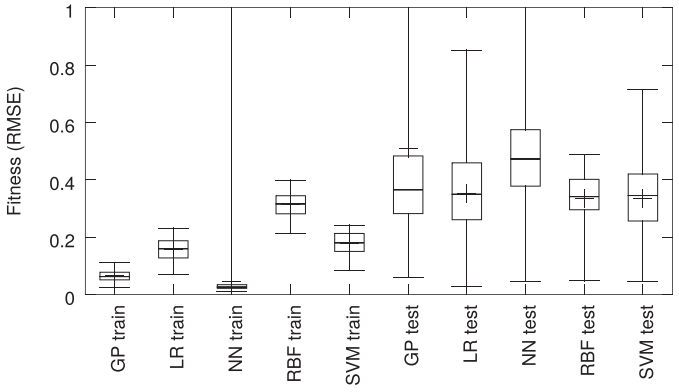

Fig. 10. The results on the IS dataset (S) for size 30.

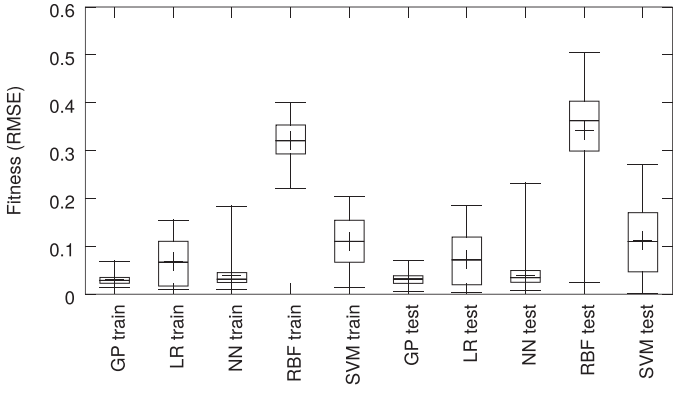

Fig. 11. The results on the IS dataset (S+U) for size 30.

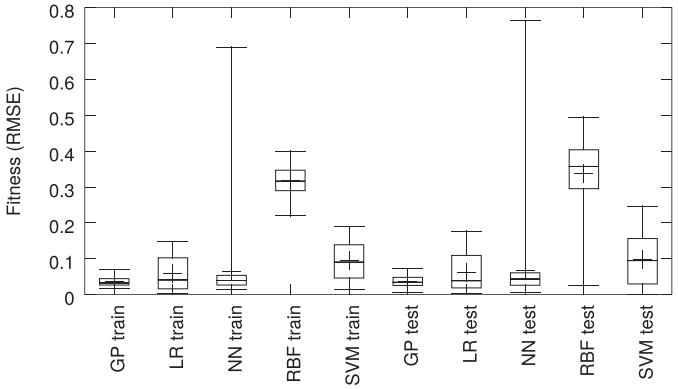

Fig. 12. The results on the IS dataset (S+U+T) for size 30.

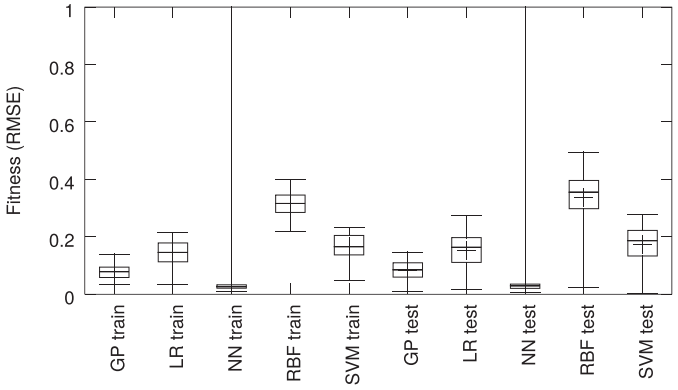

Fig. 13. The results on the IS dataset (S+T) for size 30.

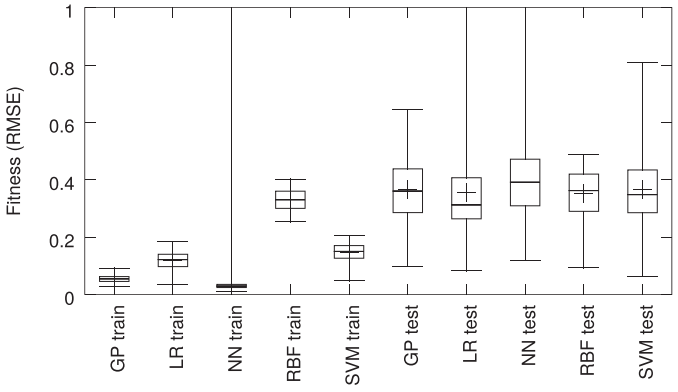

Fig. 14. The results on the IS dataset (S) for size 40.

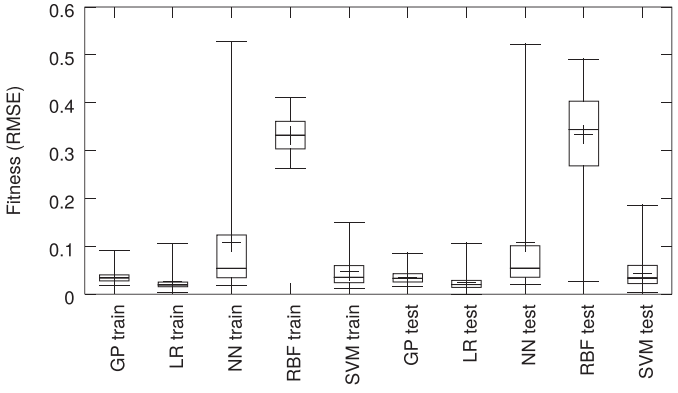

Fig. 15. The results on the IS dataset (S+U) for size 40.

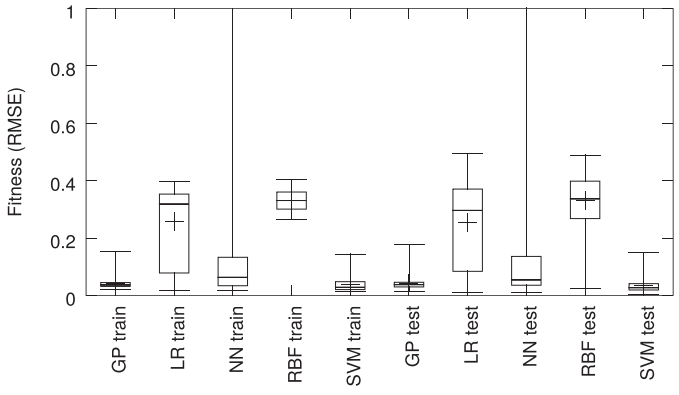

Fig. 16. The results on the IS dataset (S+U+T) for size 40.

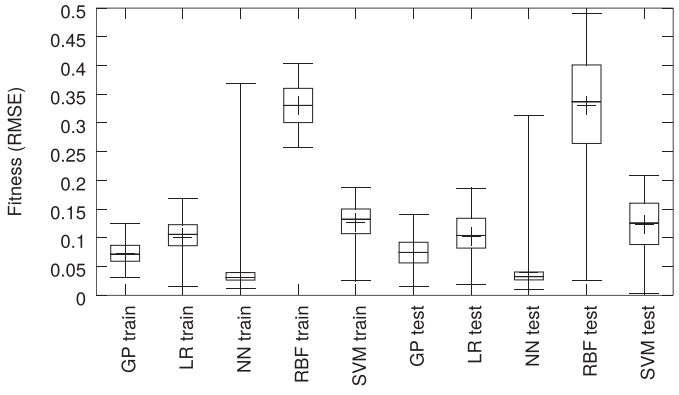

Fig. 17. The results on the IS dataset (S+T) for size 40.

considered. The target we want to predict is the average score of the remaining reviews. We use the following notation:

- (S): in this case we considered, for prediction purposes, only the review scores;
- (S+U): in this case, we considered the review scores and their usefulness;
- (S+T): in this case, we considered the review scores and the time of the reviews; and
- (S+T+U): in this case, we considered the review scores, the time of the reviews and the usefulness of the reviews.

Further, in the figures, LR stands for linear regression (Weisberg, 2005), RBF stands for radial basis function network (Haykin, 1999), SVM refers to support vector machines (Schölkopf & Smola, 2002) and NN refers to feed-forward artificial neural networks trained with the Backpropagation learning rule (Gurney, 1997). In all figures, the first five boxes, from left to right, represent the results obtained with the compared five methods on the training set, while the remaining five boxes refer to the results on the test set.

If we consider the results obtained in the case of 10 reviews (Figs. 2–5), GP outperforms all the other methods on both the training and the test set, with the only exception of the (S) case where GP is outperformed by LR on the test set. If we consider the case of 20 reviews (Figs. 6–9), GP outperforms all the other methods for the (S+U) case on both the training and test sets. For the (S+U+T) case, GP and NN are the methods that gave the best performance on both the training and test sets. On the other hand, GP is outperformed by the other methods for the (S) and the (S+T) cases. For 30 reviews (Figs. 10–13), GP and NN are the best methods for both the training and test sets, except for the (S) case on the test set. Finally, for 40 reviews (Figs. 14–17), GP and NN are the best methods on both the training and test sets,

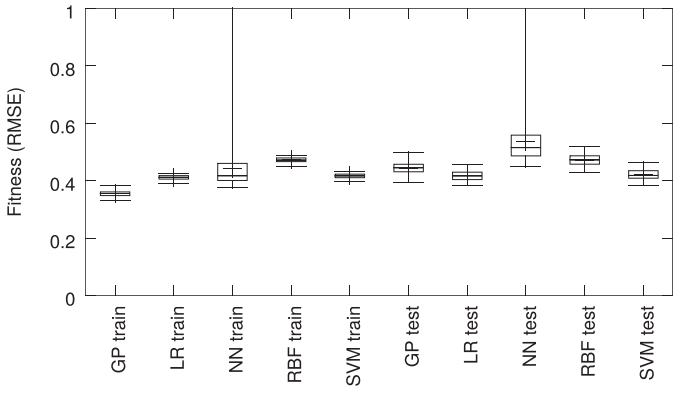

Fig. 18. The results on the KS dataset (S) for size 10.

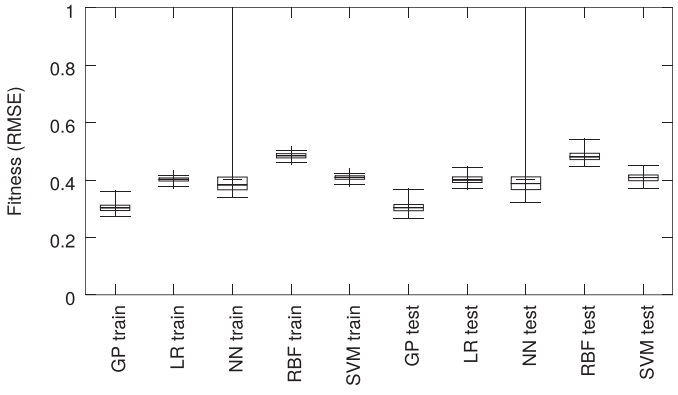

Fig. 19. The results on the KS dataset (S+U) for size 10.

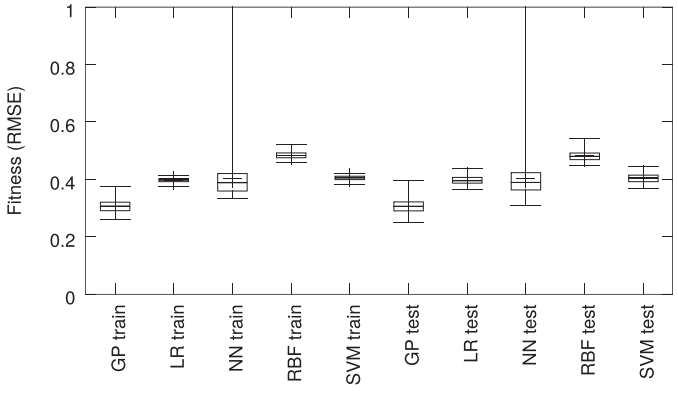

Fig. 20. The results on the KS dataset (S+U+T) for size 10.

except for the (S) case on the test set and the (S+T) case on the training and test sets. In the latter case, GP is outperformed by NN. As a partial conclusion, we can state that GP, in general, performs better than the competitors when the number of reviews is small (10 or 20). For a larger number of reviews (30 or 40), the performance of NN increases and NN have a performance that is comparable, and in some cases even better than GP. However, also for the cases of 30 and 40 reviews, GP is reliably the best, or second best, performer among the compared methods.

Plots shown in Fig. 18–33 report the results obtained for the KS dataset.

Considering 10 reviews for the KS dataset (Figs. 18–21), we can see that GP outperforms all the other methods on both the training and test sets. GP is also generally the method with the best performance for the case of 20 reviews (Figs. 22–25), but NN often have a performance comparable to it. The same thing can also be said

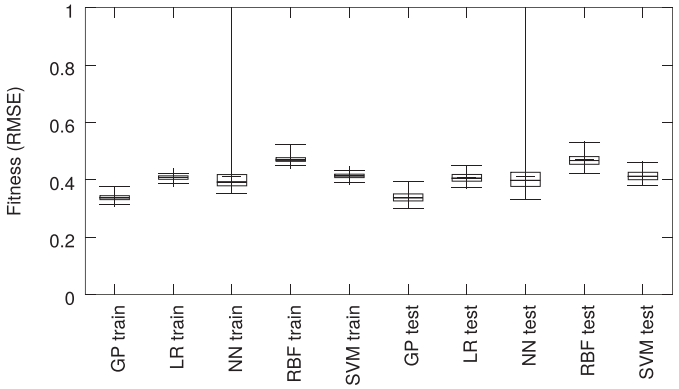

Fig. 21. The results on the KS dataset (S+T) for size 10.

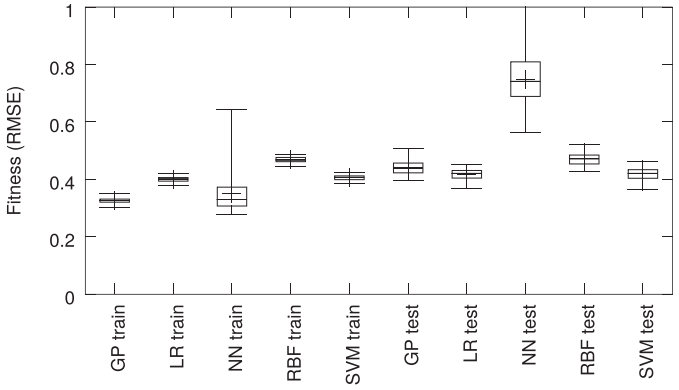

Fig. 22. The results on the KS dataset (S) for size 20.

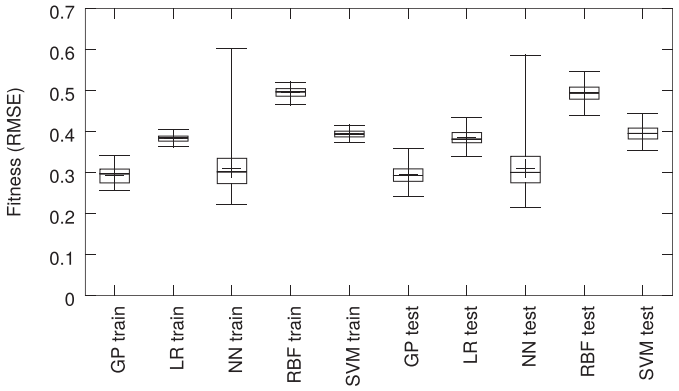

Fig. 23. The results on the KS dataset (S+U) for size 20.

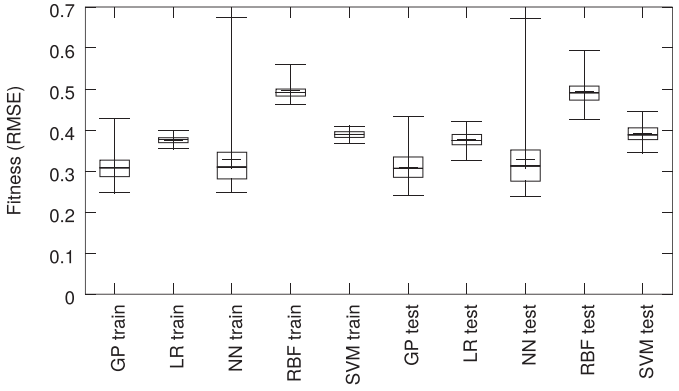

Fig. 24. The results on the KS dataset (S+U+T) for size 20.

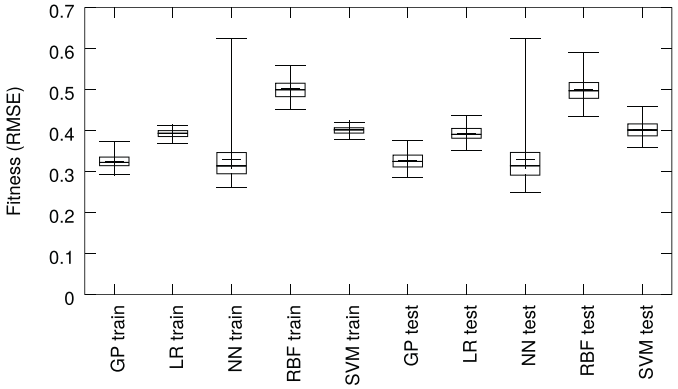

Fig. 25. The results on the KS dataset (S+T) for size 20.

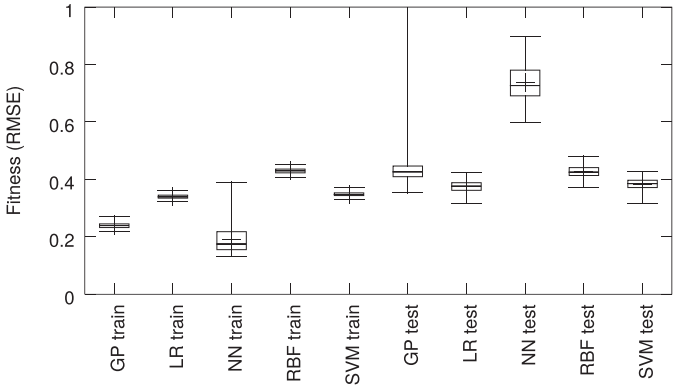

Fig. 26. The results on the KS dataset (S) for size 30.

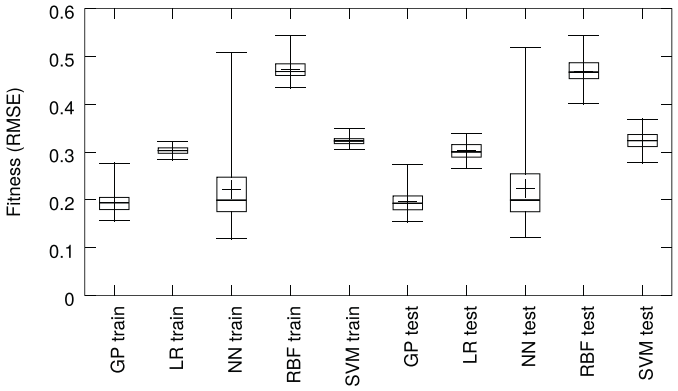

Fig. 27. The results on the KS dataset (S+U) for size 30.

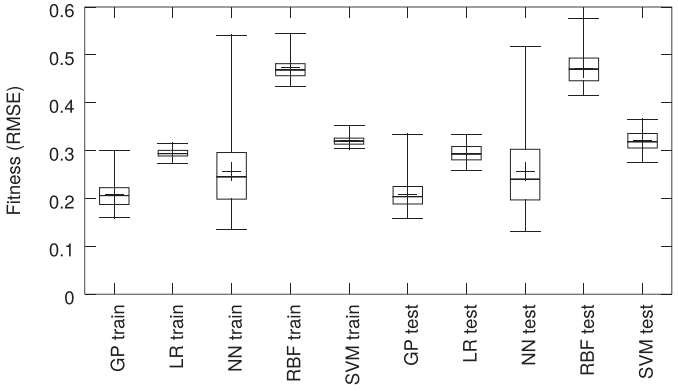

Fig. 28. The results on the KS dataset (S+U+T) for size 30.

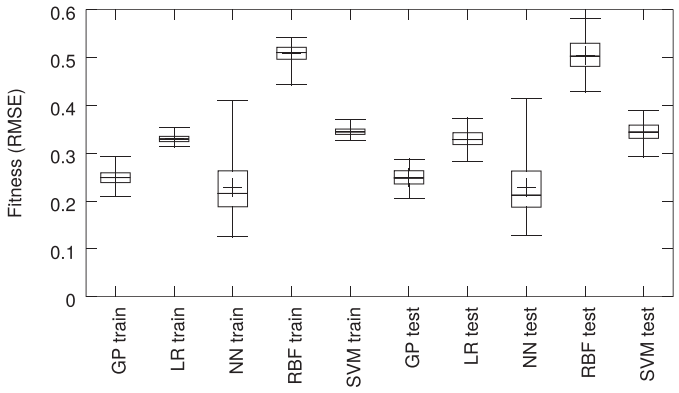

Fig. 29. The results on the KS dataset (S+T) for size 30.

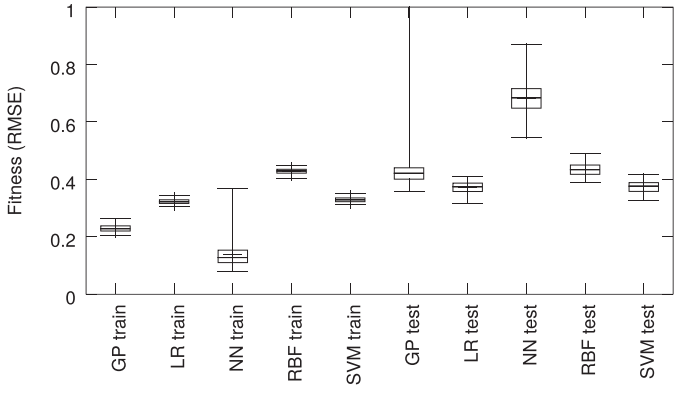

Fig. 30. The results on the KS dataset (S) for size 40.

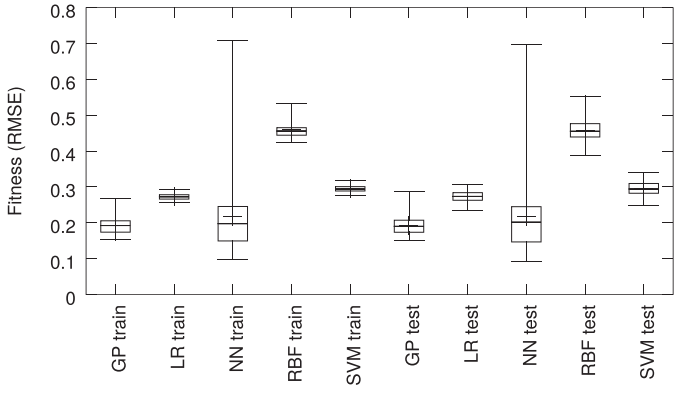

Fig. 31. The results on the KS dataset (S+U) for size 40.

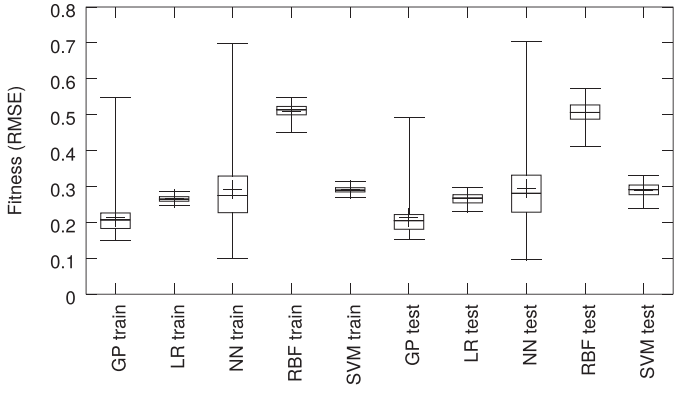

Fig. 32. The results on the KS dataset (S+U+T) for size 40.

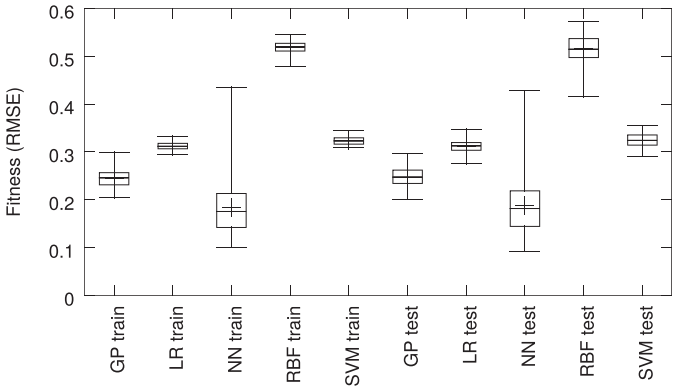

Fig. 33. The results on the KS dataset (S+T) for size 40.

for 30 reviews (Figs. 26–29), with the only exception of the (S) case where LR outperforms GP on the test set, and NN is the worst performer on the test set. Finally, for 40 reviews (Figs. 30–33), GP outperforms all the other methods on both the training and test sets, except for the (S) case in which it is outperformed by NN on the training set and LR and SVM on the test set, and the (S+T) case in which GP is outperformed by NN on both the training and test sets. All in all, we may conclude that the results obtained on the KS dataset confirm the same trend as those already discussed for the IS dataset: GP is, in general, the best method for 10 or 20 reviews, and it has a performance that is better than, or comparable to, NN (that together with GP is the best performer) for 30 and 40 reviews. These experiments clearly demonstrate the appropriateness of GP for solving the studied problem when also considering that, unlike the other techniques considered, the performance of GP is quite similar on both the training and test sets for all studied cases. Further, when looking at the boxes of GP in the previous figures and comparing them to the boxes in the other figures, we can see the results of GP are characterized with low variance. All of these aspects make GP an appropriate technique for understanding e-commerce clients' preferences when using a set of reviews.

To analyze the statistical significance of the results discussed so far, a set of tests was performed on the median errors. The Wilcoxon rank-sum test for pairwise data comparison, a rank-based statistic, was used under the alternative hypothesis that the samples from the first set are smaller or equal to the values of the second sample with a probability exceeding 0.5. A confidence value of $\alpha = 0.1$ was used and, considering the presence of more than two samples, a Bonferroni correction for this value was applied. The p-values obtained are reported from Tables 1–8.

Considering the IS dataset (Tables 1–4), it is possible to observe that for the training set GP produces results that are statistically better than those produced by the other techniques on a large set of configurations. Only for 20 reviews do GP and NN perform comparably for configurations (S), (S+T), and (S+T+U). Regarding the performance on the test set, it is possible to draw similar considerations: GP is the best performer in the large majority of configurations, with the only exceptions of the following cases: For the (S) configuration, GP, LR, RBF, and SVM produce results that are not statistically different, while for the (S+T) and (S+T+U) configurations with 20 reviews GP and NN produce results that are comparable (i.e. the difference in terms of median RMSE is not statistically significant).

On the KS dataset (Tables 5–8), GP outperforms all the other methods on the training set and, in a large number of cases, the differences between GP and the other methods are statistically significant. Only for the case of 20 reviews do GP and NN perform comparably for the (S), (S+T), and (S+T+U) configuration. This

Table 1
p-value returned by the Mann–Whitney test on the IS dataset for size 10.

	LR (train)	NN (train)	RBF (train)	SMO (train)	LR (test)	NN (test)	RBF (test)	SMO (test)
S	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.6059	1.0000
S+U	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
S+T	0.0000	0.0001	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000
S+T+U	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 2
p-value returned by the Mann–Whitney test on the IS dataset for size 20.

	LR (train)	NN (train)	RBF (train)	SMO (train)	LR (test)	NN (test)	RBF (test)	SMO (test)
S	0.0000	1.0000	0.0000	0.0000	1.0000	0.0000	0.5515	1.0000
S+U	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
S+T	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
S+T+U	0.0000	0.9353	0.0000	0.0000	0.0000	0.9433	0.0000	0.0000

Table 3
p-value returned by the Mann–Whitney test on the IS dataset for size 30.

	LR (train)	NN (train)	RBF (train)	SMO (train)	LR (test)	NN (test)	RBF (test)	SMO (test)
S	0.0000	1.0000	0.0000	0.0000	0.8605	0.0001	0.9809	0.9690
S+U	0.0001	0.0088	0.0000	0.0000	0.0001	0.0104	0.0000	0.0000
S+T	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
S+T+U	0.1915	0.0183	0.0000	0.0000	0.1073	0.0073	0.0000	0.0000

Table 4
p-value returned by the Mann–Whitney test on the IS dataset for size 40.

	LR (train)	NN (train)	RBF (train)	SMO (train)	LR (test)	NN (test)	RBF (test)	SMO (test)
S	0.0000	1.0000	0.0000	0.0000	0.9733	0.0631	0.7538	0.5861
S+U	1.0000	0.0000	0.0000	0.0789	1.0000	0.0000	0.0000	0.3048
S+T	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
S+T+U	0.0000	0.0000	0.0000	0.9999	0.0000	0.0000	0.0000	1.0000

Table 5
p-value returned by the Mann–Whitney test on the KS dataset for size 10.

	LR (train)	NN (train)	RBF (train)	SMO (train)	LR (test)	NN (test)	RBF (test)	SMO (test)
S	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000
S+U	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
S+T	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
S+T+U	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 6
p-value returned by the Mann-Whitney test on the KS dataset for size 20.

	LR (train)	NN (train)	RBF (train)	SMO (train)	LR (test)	NN (test)	RBF (test)	SMO (test)
S	0.0000	0.1942	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000
S+U	0.0000	0.0570	0.0000	0.0000	0.0000	0.0401	0.0000	0.0000
S+T	0.0000	0.9857	0.0000	0.0000	0.0000	0.9885	0.0000	0.0000
S+T+U	0.0000	0.1989	0.0000	0.0000	0.0000	0.2093	0.0000	0.0000

Table 7
p-value returned by the Mann–Whitney test on the KS dataset for size 30.

	LR (train)	NN (train)	RBF (train)	SMO (train)	LR (test)	NN (test)	RBF (test)	SMO (test)
S	0.0000	1.0000	0.0000	0.0000	1.0000	0.0000	0.4557	1.0000
S+U	0.0000	0.0368	0.0000	0.0000	0.0000	0.0279	0.0000	0.0000
S+T	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
S+T+U	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 8
p-value returned by the Mann–Whitney test on the KS dataset for size 40.

	LR (train)	NN (train)	RBF (train)	SMO (train)	LR (test)	NN (test)	RBF (test)	SMO (test)
S	0.0000	1.0000	0.0000	0.0000	1.0000	0.0000	0.0001	1.0000
S+U	0.0000	0.2207	0.0000	0.0000	0.0000	0.1428	0.0000	0.0000
S+T	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
S+T+U	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

behavior is very similar to that observed for the IS dataset. Analyzing the performance on the test set, it is possible to see that GP has a comparable performance to LR and SVM for the (S) configuration. Also, GP and NN produce results that are not statistically different for the case of 20 reviews in the (S+T), (S+U), and (S+T+U) configurations.

In conclusion, the results of the statistical tests confirm the suitability of the proposed GP-based technique for addressing the problem at hand. As a final consideration, it is important to point out that it is not only the reviews' score that is important for predicting a product's success, but also the usefulness of a review is particularly relevant and may contribute to a better prediction. Hence, e-commerce managers who provide clients with the opportunity to rate a product and to also rate existing reviews should have a competitive advantage with respect to e-commerce stores that do not provide these features.

5. Conclusions

We proposed a genetic programming system for predicting review scores based on a subset of existing reviews. The proposed system uses genetic operators that are able to integrate semantic awareness into the search process. The use of these operators induces a unimodal fitness landscape in every problem that entails finding a match between predicted values and targets (like regression and classification problems). Considering the particular problem under scrutiny, it is possible to draw some interesting conclusions: in both datasets considered a small subset of all existing reviews (review scores and other attributes) was sufficient for predicting the average review scores better than using, as a predictor, the average of the known review scores. This is a very important point for business: in electronic commerce a large amount of data is available and the knowledge extraction process can be a very time-consuming task. Having a system available that is able to guarantee good predictive accuracy can speed up the entire process. More specifically, the best prediction was achieved on both the training and test sets, considering the review scores and their usefulness. Hence, while review score is an important attribute, the usefulness of the review (which may be seen as a measure of the review's quality) also plays a primary role in achieving good predictive accuracy. In a more general perspective, this study offers e-commerce managers a tool for more comprehensively understanding customer behavior with regard to new and repeated purchases. We hope this study paves the way for future research in the area.

References

Anderson, E. W., Fornell, C., & Lehmann, D. R. (1994). Customer satisfaction, market share, and profitability: Findings from sweden. *Journal of Marketing, 58*(3), pp.53–66.

Balasubramanian, S., Konana, P., & Menon, N. M. (2003). Customer satisfaction in virtual environments: A study of online investing. *Management Science, 49*(7), 871–889.

Bazaarvoice (2007). Social commerce report 2007. http://www.bazaarvoice.com/.

Beadle, L., & Johnson, C. G. (2009). In A. Tyrrell (Ed.), *Semantically driven mutation in genetic programming* (pp. 1336–1342). Trondheim, Norway: 2009 ieee congress on evolutionary computation.

Bendoly, E., & Kaefer, F. (2004). Business technology complementarities: Impacts of the presence and strategic timing of ERP on B2B e-commerce technology efficiencies. *Omega, 32*(5), 395–405.

Bergendahl, G. (2005). Models for investment in electronic commerce: financial perspectives with empirical evidence. *Omega, 33*(4), 363–376.

Castelli, M., Silva, S., & Vanneschi, L. (2014a). A c++ framework for geometric semantic genetic programming. *Genetic Programming and Evolvable Machines*, 1–9. doi:10.1007/s10710-014-9218-0.

Castelli, M., Vanneschi, L., & Silva, S. (2014b). Semantic search-based genetic programming and the effect of intron deletion. *Cybernetics, IEEE Transactions on, 44*(1), 103–113.

Cui, G., Lui, H.-K., & Guo, X. (2012). The effect of online consumer reviews on new product sales. *International Journal of Electronics and Commerce, 17*(1), 39–58.

Ganu, G., Elhadad, N., & Marian, A. (2009). Beyond the stars: Improving rating predictions using review text content. In *Webdb: 9* (pp. 1–6).

Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega, 28*(6), 725–737.

Gupta, N., Di Fabbrizio, G., & Haffner, P. (2010). Capturing the stars: predicting ratings for service and product reviews. In *Proceedings of the naacl hlt 2010 workshop on semantic search* (pp. 36–43). Association for Computational Linguistics.

Gurney, K. (1997). An introduction to neural networks. *An Introduction to Neural Networks*. Taylor & Francis.

Haykin, S. (1999). *Neural networks: A comprehensive foundation*. Prentice Hall.

Hu, N., Liu, L., & Sambamurthy, V. (2011). Fraud detection in online consumer reviews. *Decision Support Systems, 50*(3), 614–626. On quantitative methods for detection of financial fraud.

Kim, C., Galliers, R. D., Shin, N., Ryoo, J.-H., & Kim, J. (2012). Factors influencing internet shopping value and customer repurchase intention. *Electronic Commerce Research and Applications, 11*(4), 374–387.

Koza, J. R. (1992). *Genetic programming: On the programming of computers by means of natural selection*. Cambridge, MA, USA: MIT Press.

Koza, J. R. (2010). Human-competitive results produced by genetic programming. *Genetic Programming and Evolvable Machines, 11*(3-4), 251–284.

Lackermair, G., Kailer, D., & Kanmaz, K. (2013). Importance of online product reviews from a consumer's perspective. *Advances in Economics and Business, 1*(1), 1–5.

Lee, H., Han, J., & Suh, Y. (2014). Gift or threat? An examination of voice of the customer: The case of mystarbucksidea.com. *Electronic Commerce Research and Applications, 13*(3), 205–219.

Liu, C., & Arnett, K. P. (2000). Exploring the factors associated with web site success in the context of electronic commerce. *Information & Management, 38*(1), 23–33.

McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. In *In proceedings of the 7th ACM conference on recommender systems, RecSys '13* (pp. 165–172). New York, NY, USA: A.C.M.

Moraglio, A., Krawiec, K., & Johnson, C. G. (2012). Geometric semantic genetic programming. In C. A. Coello Coello, V. Cutello, K. Deb, S. Forrest, G. Nicosia, & M. Pavone (Eds.), *Parallel problem solving from nature, ppsn xii (part 1)*. In *Lecture Notes in Computer Science: 7491* (pp. 21–31). Springer.

Nielsen (2010). Global online shopping report. http://acnielsen.co.th/us/en/newswire/2010/global-online-shopping-report.html.

Qiu, J., Lin, Z., & Li, Y. (2015). Predicting customer purchase behavior in the e-commerce context. *Electronic Commerce Research*, 1–26.

Qu, L., Ifrim, G., & Weikum, G. (2010). The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 913–921). Association for Computational Linguistics.

Schölkopf, B., & Smola, A. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. *Adaptive computation and machine learning*. MIT Press.

Stadler, P. (1995). Towards a theory of landscapes. In R. López-Pena, H. Waelbroeck, R. Capovilla, R. García-Pelayo, & F. Zertuche (Eds.), *Complex systems and binary networks*. In *Lecture Notes in Physics: 461-461* (pp. 78–163). Springer Berlin / Heidelberg.

Szymanski, D., & Henard, D. (2001). Customer satisfaction: A meta-analysis of the empirical evidence. *Journal of the Academy of Marketing Science, 29*(1), 16–35.

Teo, T. S., & Liu, J. (2007). Consumer trust in e-commerce in the united states, singapore and china. *Omega, 35*(1), 22–38.

Vanneschi, L. (2017). In O. Schütze, L. Trujillo, P. Legrand, & Y. Maldonado (Eds.), *An introduction to geometric semantic genetic programming* (pp. 3–42)). Cham: Springer International Publishing.

Vanneschi, L., Castelli, M., Manzoni, L., & Silva, S. (2013). A new implementation of geometric semantic GP and its application to problems in pharmacokinetics. In K. Krawiec, A. Moraglio, T. Hu, A. Etaner-Uyar, & B. Hu (Eds.), *Genetic programming*. In *Lecture Notes in Computer Science: 7831* (pp. 205–216). Springer Berlin Heidelberg.

Vanneschi, L., Castelli, M., & Silva, S. (2014a). A survey of semantic methods in genetic programming. *Genetic Programming and Evolvable Machines, 15*(2), 195–214.

Vanneschi, L., Silva, S., Castelli, M., & Manzoni, L. (2014b). Geometric semantic genetic programming for real life applications. In *Genetic programming theory and practice xi* (pp. 191–209). Springer New York.

Verbraken, T., Goethals, F., Verbeke, W., & Baesens, B. (2014). Predicting online channel acceptance with social network data. *Decision Support Systems, 63*(0), 104–114.

Weisberg, S. (2005). Applied linear regression. *Wiley Series in Probability and Statistics*. Wiley.

Weka machine learning project (2013). Weka. http://www.cs.waikato.ac.nz/~ml/weka.