# Applying machine learning to the problem of choosing a heuristic to select the variable ordering for cylindrical algebraic decomposition

Huang, Z., England, M., Wilson, D., Davenport, J. H., Paulson, L. C. & Bridge, J.

**Author post-print (accepted) deposited by Coventry University's Repository**

# Applying machine learning to the problem of choosing a heuristic to select the variable ordering for cylindrical algebraic decomposition

Zongyan Huang[1], Matthew England[2], David Wilson[2],
James H. Davenport[2], Lawrence C. Paulson[1] and James Bridge[1]

[1] University of Cambridge Computer Laboratory, Cambridge CB3 0FD, U.K.
[2] University of Bath, Department of Computer Science, Bath, BA2 7AY, U.K.
{zh242, lp15, jpb65}@cam.ac.uk,
{J.H.Davenport, M.England, D.J.Wilson}@bath.ac.uk

**Abstract.** Cylindrical algebraic decomposition(CAD) is a key tool in computational algebraic geometry, particularly for quantifier elimination over real-closed fields. When using CAD, there is often a choice for the ordering placed on the variables. This can be important, with some problems infeasible with one variable ordering but easy with another. Machine learning is the process of fitting a computer model to a complex function based on properties learned from measured data. In this paper we use machine learning (specifically a support vector machine) to select between heuristics for choosing a variable ordering, outperforming each of the separate heuristics.

**Keywords:** machine learning, support vector machine, symbolic computation, cylindrical algebraic decomposition, problem formulation

## 1 Introduction

Cylindrical algebraic decomposition (CAD) is a key tool in real algebraic geometry. It was first introduced by Collins [18] to implement quantifier elimination over the reals, but has since been applied to applications including robot motion planning [49], programming with complex valued functions [22], optimisation [28] and epidemic modelling [15]. Decision methods for real closed fields are of great use in theorem proving [25]. METITARSKI [1], for example, decides the truth of statements about special functions using CAD and rational function bounds.

When using CAD, we often have a choice over which variable ordering to use. It is well known that this choice is very important and can dramatically affect the feasibility of a problem. In fact, Brown and Davenport [14] presented a class of problems in which one variable ordering gave output of double exponential complexity in the number of variables and another output of a constant size. Heuristics have been developed to help with this choice, with Dolzmann et al. [23] giving the best known study. However, in CICM last year [8], it was shown that even the best known heuristic could be misled. Although that paper provided

an alternative heuristic, this had its own shortcomings, and it now seems likely that no one heuristic is suitable for all problems.

Our thesis is that the best heuristic to use is dependent upon the problem considered. However, the relationship between the problems and heuristics is far from obvious and so we investigate whether machine learning can help with these choices. Machine learning is a branch of artificial intelligence. It uses statistical methods to infer information from supplied data which is then used to make predictions for previously unseen data [2]. We have applied machine learning (specifically a support vector machine) to the problem of selecting a variable ordering for both CAD itself and quantifier elimination by CAD, using the nlsat dataset [50] of fully existentially quantified problems. Our results show that the choices made by machine learning are on average superior to both any individual heuristic and to picking a heuristic at random. The results also provide some new insight on the heuristics themselves. This appears to be the first application of machine learning to problem formulation for computer algebra, although it follows recent application to theorem proving [10, 31].

We conclude the introduction with background theory on CAD and machine learning. Then in Sections 2, 3 and 4 we describe our experiment, its results and how they may be extended in the future. Finally in Section 5 we give our conclusions and ideas for future work.

## 1.1  Quantifier elimination and CAD

Let $Q_i \in \{\exists, \forall\}$ be quantifiers and $\phi$ be some quantifier free formula. Then given

$$\Phi(x_1, \ldots, x_k) := Q_{k+1} x_{k+1} \ldots Q_n x_n \, \phi(x_1, \ldots, x_n),$$

*quantifier elimination* (QE) is the problem of producing a quantifier free formulae $\psi(x_1, \ldots, x_k)$ equivalent to $\Phi$. In the case $k = 0$ this reduces to the *decision problem*, is $\Phi$ true? Tarski proved that QE was possible for semi-algebraic formulae (polynomials and inequalities) over $\mathbb{R}$ [47]. However, the complexity of Tarski's method is non-elementary (indescribable as a finite tower of exponentials) and so CAD was a major breakthrough when introduced, despite complexity doubly exponential in the number of variables. For some problems QE is possible through algorithms with better complexity (see for example the survey by Basu [5]), but CAD implementations remain the best general purpose approach.

Collins' algorithm [3] works in two stages. First, *projection* calculates sets of projection polynomials $S_i$ in variables $(x_1, \ldots, x_i)$. This is achieved by repeatedly applying a projection operator onto a set of polynomials, producing a set with one variable fewer. We start with the polynomials from $\phi$ and eliminate variables this way until we have the set of univariate polynomials $S_1$.

Then in the *lifting* stage, decompositions of real space in increasing dimensions are formed according to the real roots of those polynomials. First, the real line is decomposed according to the roots of the polynomials in $S_1$. Then over each cell $c$ in that decomposition, the bivariate polynomials $S_2$ are taken at a sample point and a decomposition of $c \times \mathbb{R}$ is produced according to their roots.

Taking the union gives the decomposition of $\mathbb{R}^2$ and we proceed this way to a decomposition of $\mathbb{R}^n$. The decompositions are cylindrical (projections of any two cells onto their first $i$ coordinates are either identical or disjoint) and each cell is a semi-algebraic set (described by polynomial relations). Collins' original algorithm used a projection operator which guaranteed CADs of $\mathbb{R}^n$ on which the polynomials in $\phi$ had constant sign, and thus $\phi$ constant truth value, on each cell. Hence only a single sample point from each cell needed to be tested and the equivalent quantifier free formula $\psi$ could be generated from the semi-algebraic sets defining the cells in the CAD of $\mathbb{R}^k$ for which $\Phi$ is true.

Since the publication of the original algorithm, there have been numerous improvements, optimisations and extensions of CAD (with a summary of the first 20 years given by Collins [19]). Of great importance is the improvement to the projection operator used. Hong [29] proved that a refinement of Collins' operator was sufficient and then McCallum [37] presented a further refinement which could only be used for input that was *well-oriented* and was in turn improved by Brown [11]. Further refinements are possible by removing the need for sign-invariance of polynomials while maintaining truth-invariance of a formula, with McCallum [38] presenting an operator for use when an equational constraint is present (an equation logically implied by a formula) and Bradford *et al.* [7] extending this to the case of multiple formulae. Collins and Hong [20] described Partial CAD for QE, where lifting over a cell is aborted if there already exists sufficient information to determine the truth of $\phi$ on that cell. Other recent CAD developments of particular note include the use of symbolic-numeric techniques in the lifting stage [33, 45] and the alternative to projection and lifting offered by decompositions of complex space via regular chains technology [17].

When using CAD we have to assign an ordering to the variables (the labels $i$ on the $x_i$ in the discussion above). This dictates the order in which the variables are eliminated during projection and thus the sub-spaces for which CADs are produced en route to a CAD of $\mathbb{R}^n$. For some applications this order is fixed but for others there may be a free or constrained choice. When using CAD for QE we must project quantified variables before unquantified ones. Further, the quantified variables should be projected in the order they occur, unless successive ones have the same quantifier in which case they may be swapped. The ordering can have a big effect on the output and performance of CAD [8, 14, 23].

### 1.2 Machine Learning

Machine learning [2] deals with the design of programs that can learn rules from data. This is often a very attractive alternative to manually constructing them when the underlying functional relationship is very complex. Machine learning techniques have been widely used in many fields, such as web searching [6], text categorization [42], robotics [44], expert systems [27] and many others.

Various machine learning techniques have been developed. McCulloch and Pitts [39] created the first computational model for *neural networks* called *threshold logic*. Following that, Rosenblatt [40] proposed the *perceptron* as an iterative algorithm for supervised classification of an input into one of several possible

non-binary outputs. A later development was the *decision tree* [2], which is a simple representation for classifying examples. The main idea here is to apply serial classifications which refine the output state. At the same time as the *decision tree* was being developed, the *multi-layer perceptron* [30] was explored. It is a modification of the standard linear perceptron and can distinguish data that are non-linearly separable.

In the last decade, the use of machine learning has spread rapidly following the invention of the *Support Vector Machine* (SVM) [41]. This was a development of the perceptron approach and gives a powerful and robust method for both classification and regression. *Classification* refers to the assignment of input examples into a given set of classes (the output being the class labels). *Regression* refers to a supervised pattern analysis in which the output is real-valued. The SVM technology can deal efficiently with high-dimensional data, and is flexible in modelling diverse sources of data. The standard SVM classifier takes a set of input data and predicts one of two possible classes from the input. Given a set of examples, each marked as belonging to one of two classes, an SVM training algorithm builds a model that assigns new examples into one of the classes. The examples used to fit the model are called training examples.

An important concept in the SVM theory is the use of a kernel function [43], which maps data into a high dimensional kernel-defined feature space and then separates samples in the transformed space. Kernel functions enable operations in feature space without ever computing the coordinates of the data in that space. Instead they simply compute the inner products between all pairs of data vectors. This operation is generally computationally cheaper than the explicit computation of the coordinates.

The machine learning experiment described in this paper uses SVM-LIGHT (see Joachims [34]) which is an implementation of SVMs in C. The SVM-LIGHT software consists of two programs: SVM LEARN and SVM CLASSIFY. SVM LEARN fits the model parameters based on the training data and user inputs (such as the kernel function and the parameter values). SVM CLASSIFY uses the generated model to classify new samples. It calculates a hyperplane of the $n$-dimensional transformed feature space, which is an affine subspace of dimension $n-1$ dividing the space into two corresponding to the two distinct classes. SVM CLASSIFY outputs margin values which are a measure of how far the sample is from this separating hyperplane. Hence the margins are a measure of the confidence in a correct prediction. A large margin represents high confidence in a correct prediction. The accuracy of the generated model is largely dependent on the selection of the kernel functions and parameter values.

## 2 Methodology

### 2.1 CAD implementation and heuristics

For the machine learning experiment we decided to focus on a single CAD implementation, QEPCAD [12]. We note that other CAD implementations are available, as discussed further in Section 4.

Qepcad is an interactive command line program written in C for performing **Q**uantifier **E**limination with **P**artial **CAD**. It was chosen as it is a competitive implementation of both CAD and QE that also allows the user some control and information during its execution. We used Qepcad with its default settings which implement McCallum's projection operator [37] and partial CAD [20]. It can also makes use of an equational constraint automatically (via the projection operator [38]) when one is explicit in the formula, (where *explicit* means the formula is a conjunction of the equational constraint with a sub-formula).

In the experiment we used three existing heuristics for picking a CAD variable ordering:

**Brown:** This heuristic chooses a variable ordering according to the following criteria, starting with the first and breaking ties with successive ones:
  (1) Eliminate a variable first if it has lower overall degree in the input.
  (2) Eliminate a variable first if it has lower (maximum) total degree of those terms in the input in which it occurs.
  (3) Eliminate a variable first if there is a smaller number of terms in the input which contain the variable.
  It is labelled after Brown who suggested it [13].
**sotd:** This heuristic constructs the full set of projection polynomials for each permitted ordering and selects the ordering whose corresponding set has the lowest sum of total degrees for each of the monomials in each of the polynomials. It is labelled sotd for *sum of total degree* and was suggested by Dolzmann, Seidell and Sturm [23], whose study found it to be a good heuristic for both CAD and QE by CAD.
**ndrr:** This heuristic constructs the full set of projection polynomials for each ordering and selects the ordering whose set has the lowest number of distinct real roots of the univariate polynomials within. It is labelled ndrr for *number of distinct real roots* and was suggested by Bradford *et al.* [8]. Ndrr was shown to assist with examples where sotd failed.

Brown's heuristic has the advantage of being very cheap, since it acts only on the input and checks only simple properties. The ndrr heuristic is the most expensive (requiring real root isolation), but is the only one to explicitly consider the real geometry of the problem (rather than the geometry in complex space).

All three heuristics may identify more than one variable ordering as a suitable choice. In this case we took the heuristic's choice to be the first of these after they had been ordered lexicographically. [1]

---

[1] This final choice may depend on the convention used for displaying the variable ordering. Qepcad and the notes where Brown introduces his heuristic [13] use the convention of ordering variables from left to right so that the last one is projected first. On the other hand, Maple and the papers introducing sotd and ndrr [8, 23] use the opposite convention. The heuristics were implemented in Maple and so ties were broken by picking the first lexicographically on the second convention. This corresponds to picking the first under a reverse lexicographical order under the Qepcad convention. The important point is that all three heuristics had ties broken under the same convention and so were treated fairly.

## 2.2 Problem data

Problems were taken from the nlsat dataset [50], chosen over more traditional CAD problem sets (such as Wilson *et al.* [48]) as these did not have sufficient numbers of problems for machine learning. 7001 three-variable CAD problems were extracted for our experiment. The number of variables was restricted for two reasons. First to make it feasible to test all possible variable orderings and second to avoid the possibility that QEPCAD will produce errors or warnings related to well-orientedness with the McCallum projection [37].

Two experiments were undertaken, applying machine learning to CAD itself and to QE by CAD. QE is clearly very important throughout engineering and the sciences, but increasingly CAD has been applied outside of this context, as discussed in the introduction. We performed separate experiments since for quantified problems QEPCAD can use the partial CAD techniques to stop the lifting process early if the outcome is already determined, while the full process is completed for unquantified ones and the two outputs can be quite different.

The problems from the nlsat dataset are all fully existential (satisfiability or SAT problems). A second set of problems for the quantifier free experiment was obtained by simply removing all quantifiers. An example of the QEPCAD input for a SAT problem is given in Figure 1 with the corresponding input for the unquantified problem in Figure 2. Of course, for such quantified problems there are better alternatives to building a CAD (see for example the work of Jovanovic and de Moura [36]). However, our decision to use only SAT problems was based on availability of data rather than it being a requirement of the technology, and so we focus on CAD only here and discuss how we might generalise our data in Section 4. For both experiments, the problems were randomly split into training sets (3545 problems in each), validation sets (1735 problems in each) and test sets (1721 problems in each) [2].

## 2.3 Evaluating the heuristics

Since each problem has three-variables and all the quantifiers are the same, all six possible variable orderings are admissible. For each ordering we had QEPCAD build a CAD and measured the number of cells. The best ordering was defined as the one resulting in the smallest cell count, (and if more than one ordering gives the minimal both orderings are considered the best). The decision to focus on cell counts (rather than say computation time) was made so that our experiment could validate the use of machine learning to CAD theory, rather than just the QEPCAD implementation. Further, it is usually the case that cell counts and timings are strongly correlated.

The heuristics (Brown, sotd and ndrr) have been implemented in MAPLE (as part of the freely available `ProjectionCAD` package [26]) and for each problem the orderings suggested by the heuristics were recorded and compared to the cell

---

[2] The data is available at `http://www.cl.cam.ac.uk/∼zh242/data`.

**Fig. 1:** Sample QEPCAD input for a quantified problem.

```
(x0,x1,x2)
0
(Ex0)(Ex1)(Ex2)[[((x0 x0) + ((x1 x1) + (x2 x2))) = 1]].
go
go
go
d-stat
go
finish
```

**Fig. 2:** Sample QEPCAD input for a quantifier free problem.

```
(x0,x1,x2)
3
[[((x0 x0) + ((x1 x1) + (x2 x2))) = 1]].
go
go
d-proj-factors
d-proj-polynomials
go
d-fpc-stat
go
```

counts produced by QEPCAD [3]. Note that all three heuristics do not discriminate on the structure of any quantifiers. As discussed above, some heuristics are more expensive than others. However, since none of the costs were prohibitive for our data set they are not considered here.

Machine learning was applied to predict which of the three heuristics will give an *optimal* variable ordering for a given problem, where *optimal* means the lowest cell count of the selected CADs. Note that in the quantified case QEPCAD can collapse stacks when sufficient truth values for the constituent cells have been discovered to determine a truth value for the base cell. Hence, since our problems are all fully existential, the output for all quantified problems is always a single cell: true or false. Therefore, in these cases it was not the number of cells in the output that was used but instead the number of cells constructed during the process (hence the statistics commands in Figures 1 and 2 differ).

### 2.4 Problem features

To apply machine learning, we need to identify features of the CAD problems that might be relevant to the correct choice of the heuristics. A feature is an

---

[3] When comparing care must be taken when changing between the different variable ordering conventions (see Footnote 1).

aspect or measure of the problem that may be expressed numerically. Table 1 shows the 11 features that we identified, where $(x_0, x_1, x_2)$ are the three variable labels used in all our problems. The number of features is quite small, compared to other machine learning experiments. They were chosen as easily computable features of the problems which could affect the performances of the heuristics. Other features were considered (such as the maximum coefficient and the proportion of constraints that were equations) but were not found to be useful. Further investigation into feature selection may be a topic of our future work.

**Table 1:** Description of the features used. The proportion of a variable occurring in polynomials is the number of polynomials containing the variable divided by total number of polynomials. The proportion of a variable occurring in monomials is the number of terms containing the variable divided by total number of terms in polynomials.

| Feature number | Description |
| --- | --- |
| 1 | Number of polynomials. |
| 2 | Maximum total degree of polynomials. |
| 3 | Maximum degree of $x_0$ among all polynomials. |
| 4 | Maximum degree of $x_1$ among all polynomials. |
| 5 | Maximum degree of $x_2$ among all polynomials. |
| 6 | Proportion of $x_0$ occurring in polynomials. |
| 7 | Proportion of $x_1$ occurring in polynomials. |
| 8 | Proportion of $x_2$ occurring in polynomials. |
| 9 | Proportion of $x_0$ occurring in monomials. |
| 10 | Proportion of $x_1$ occurring in monomials. |
| 11 | Proportion of $x_2$ occurring in monomials. |

Each feature vector in the training set was associated with a label, $+1$ (positive examples) or $-1$ (negative examples), indicating in which of two classes it was placed. To take Brown's heuristic as an example, a corresponding training set was derived with each problem labelled $+1$ if Brown's heuristic suggested a variable ordering with the lowest number of cells, or $-1$ otherwise.

The features could all be easily calculated from the problem input using MAPLE. For example. if the input formula is defined using the set of polynomials

$$\{-6x_0^2 - x_2^3 - 1, \quad x_0^4 x_2 + 9x_1, \quad x_0 + x_0^2 - x_2 x_0 - 5\}$$

then the problem will have the feature vector

$$\left[3, 5, 4, 1, 3, 1, \frac{1}{3}, 1, \frac{5}{9}, \frac{1}{9}, \frac{1}{3}\right].$$

After the feature generation process, the training data (feature vectors) were normalized so that each feature had zero mean and unit variance across the set. The same normalization was then also applied to the validation and test sets.

## 2.5 Parameter Optimization

SVM-LIGHT was used to do the classification for this experiment. As stated in Section 1.2, SVMs use kernel functions to map the data into higher dimensional spaces where the data may be more easily separated. SVM-LIGHT has four standard kernel functions: linear, polynomial, sigmoid tanh and radial basis function. For each kernel function, there are associated parameters which must be set. An earlier experiments applying machine learning to an automated theorem prover [9] found the radial basis function (RBF) kernel performed well in finding a relation between the simple algebraic features and the best heuristic choice. Hence the same kernel was selected for this experiment (other kernel functions may be tested in future work). The RBF function is defined as:

$$K(x, x\prime) = \exp\left(-\gamma ||x - x\prime||^2\right)$$

where $K$ is the kernel function, $x$ and $x\prime$ are feature vectors. There is a single parameter $\gamma$ in the RBF kernel function. Besides the parameter $\gamma$, two other parameters are involved in the SVM fitting process. The parameter $C$ governs the trade-off between margin and training error, and the cost factor $j$ is used to correct imbalance in the training set and we set it equal to the ratio between negative and positive samples. Given a training set, we can easily compute the value of parameter $j$ by looking at the sign of the samples. However, it is not that trivial to find the optimal values of $\gamma$ and $C$.

In machine learning, *Matthew's correlation coefficient* (MCC) [4] is often used to evaluate the performance of the binary classifications. It takes into account true and false positives and negatives:

$$\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

In this equation, TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. The denominator is set to 1 if any sum term is zero. This measure has the value 1 if perfect prediction is attained, 0 if the classifier is performing as a random classifier, and $-1$ if the classifier exactly disagrees with the data.

A grid-search optimisation procedure was used with the training and validation set, involving a search over a range of $(\gamma, C)$ values to find the pair which would maximize MCC. We tested a commonly used range of value of $\gamma$ (varied between $2^{-15}, 2^{-14}, 2^{-13}, \ldots, 2^3$) and $C$ (varied between $2^{-5}, 2^{-4}, 2^{-3}, \ldots, 2^{15}$) in our grid search process [32]. Following the completion of the grid-search, the values for kernel function and model parameters giving optimal MCC results were selected for each individual CAD heuristic classifier. We also performed a similar calculation, selecting parameters to maximise the $F_1$-score [35], but the results using MCC were superior.

The classifiers with optimal $(\gamma, C)$ were applied to the test set to output the margin values [21]. In an ideal case, only one classifier would return a positive result for any problem, where selecting a best heuristic is just a case of observing

which classifier returns a positive result. However, in practice, more than one classifier will return a positive result for some problems, while no classifiers may return a positive for others. Thus, instead we used the relative magnitudes of the classifiers in our experiment. The classifier with most positive (or least negative) margin was selected to indicate the best decision procedure for the selection.

## 3 Results

The experiment was run as described in Section 2. We use the number of problems for which a selected variable ordering is optimal to measure the efficacy of each heuristic separately, and of the heuristic selected by machine learning.

Table 2 breaks down the results into a set of mutually exclusive outcomes that describe all possibilities. The column headed 'Machine Learning' indicates the heuristic selected by the machine learned model with the next three columns indicating each of the fixed heuristics tested. For each of these four heuristics, we may ask the question "Did this heuristic select the optimal variable ordering?" A 'Y' in the table indicates yes and an 'N' indicates no, with each of the 13 cases listed covering all possibilities. Note that at least one of the fixed heuristics must have a 'Y' since, by definition, the optimal ordering is obtained by at least one heuristic while if they all have a Y it is not possible for machine learning to fail. For each of these cases we list the number of problems for which this case occurred for both the quantifier free and quantified experiments.

**Table 2:** Categorising the problems into a set of mutually exclusive cases characterised by which heuristics were successful.

| Case | Machine Learning | sotd | ndrr | Brown | Quantifier Free | Quantified |
|------|------------------|------|------|-------|-----------------|------------|
| 1    | Y                | Y    | Y    | Y     | 399             | 573        |
| 2    | Y                | Y    | Y    | N     | 146             | 96         |
| 3    | N                | Y    | Y    | N     | 39              | 24         |
| 4    | Y                | Y    | N    | Y     | 208             | 232        |
| 5    | N                | Y    | N    | Y     | 35              | 43         |
| 6    | Y                | N    | Y    | Y     | 64              | 57         |
| 7    | N                | N    | Y    | Y     | 7               | 11         |
| 8    | Y                | Y    | N    | N     | 106             | 66         |
| 9    | N                | Y    | N    | N     | 106             | 75         |
| 10   | Y                | N    | Y    | N     | 159             | 101        |
| 11   | N                | N    | Y    | N     | 58              | 89         |
| 12   | Y                | N    | N    | Y     | 230             | 208        |
| 13   | N                | N    | N    | Y     | 164             | 146        |

For many problems more than one heuristic selects the optimal variable ordering and the probability of a randomly selected heuristic giving the optimal ordering depends on how many pick it. For example, a random selection would be successful 1/3 of the time if one heuristic gives the optimal ordering or 2/3 of the time if two heuristics do so.

In Table 2, case 1 is where machine learning cannot make any difference as all heuristics are equally optimal. We compare the remaining cases pairwise. For each pair, the behaviour of the fixed heuristics are identical and the difference is whether or not machine learning picked a winning heuristic (one of the ones with a Y). We see that in each case machine learning succeeds far more often than fails. For each pair we can compare with a random heuristic selection. For example, consider cases 2 and 3 where sotd and ndrr are successful heuristics and Brown is not. A random selection would be successful 2/3 of the time. For the quantifier free examples, machine learned selection is successful $146/(146 + 39)$ or approximately 79% of the time, which is significantly better.

We repeated this calculation for the quantified case and the other pairs, as shown in Table 3. In each case the values have been compared to the chance of success when picking a random heuristic, and so there are two distinct sets in Table 3: those where only one heuristic was optimal and those where two are. We see that machine learning did better for some classes of problems than others. For example in quantifier free examples, when only one heuristic is optimal machine learning does considerably better if that one is ndrr, while if only one is not optimal machine learning does worse if is Brown. Nevertheless, the machine learning selection is better than random in every case in both experiments.

**Table 3:** Proportion of examples where machine learning picks a successful heuristic.

| sotd | ndrr | Brown | Quantifier Free | Quantified |
|------|------|-------|-----------------|------------|
| Y | Y | N | 79% (>67%) | 80% (>67%) |
| Y | N | Y | 86% (>67%) | 84% (>67%) |
| N | Y | Y | 90% (>67%) | 84% (>67%) |
| Y | N | N | 50% (>33%) | 47% (>33%) |
| N | Y | N | 73% (>33%) | 53% (>33%) |
| N | N | Y | 58% (>33%) | 59% (>33%) |

By summing the numbers in Table 2 in which Y appears in a row for the machine learned selection and each individual heuristic, we get Table 4. This compares, for both the quantifier free and quantified problem sets, the learned selection with each of the CAD heuristics on their own.

Of the three heuristics, Brown seems to be the best, albeit by a small margin. Its performance is a little surprising, both because the Brown heuristic is not so well known (having never been formally published) and because it requires little computation (taking only simple measurements on the input).

**Table 4:** Total number of problems for which each heuristic picks the best ordering.

|                 | Machine Learning | sotd | ndrr | Brown |
|-----------------|:----------------:|:----:|:----:|:-----:|
| Quantifier free | 1312             | 1039 | 872  | 1107  |
| Quantified      | 1333             | 1109 | 951  | 1270  |

For the quantifier free problems there were 399 problems where every heuristic picked the optimal, 499 where two did and 823 where one did. Hence for this problem set the chances of picking a successful heuristic at random is

$$\frac{100}{1721}\left(399 + 499 * \tfrac{2}{3} + 823 * \tfrac{1}{3}\right) \simeq 58\%$$

which compares with $100 * 1312/1721 \simeq 76\%$ for machine learning. For the quantified problems the figures are 64% and 77%. Hence machine learning performs significantly better than a random choice in both cases. Further, if we were to use only the heuristic that performed the best on this data, the Brown heuristic, then we would pick a successful ordering for approximately 64% of the quantifier free problems and 74% of the quantified problems. So we see that a machine learned choice is also superior to using any one heuristic.

## 4 Possibilities for extending the experiment

Although a large data set of real world problems was used, we note that in some ways the data was quite uniform. A key area of future work is experimentation on a wider data set to see if these results, both the benefit of machine learning and the superiority of Brown's heuristic, are verified more generally. An initial extension would be to relax the parameters used to select problems from the nlsat dataset, for example by allowing problems with more variables.

One key restriction with this dataset is that all problems have one block of existential quantifiers. Note that our restriction to this case followed the availability of data rather than any technical limitation of the machine learning. Possible ways to generalise the data include randomly applying quantifiers to the the existing problems, or randomly generating whole problems. However, this would mean the problems no longer originate from real applications, and it has been noted in the past that random problems for CAD can be unrepresentative.

We do not suggest SVM as the only suitable machine learning method for this experiment, but overall a SVM with the RBF kernel worked well here. It would be interesting to see if other machine learning methods could offer similar or even better selections. Further improvements may also come from more work on the feature selection. The features used here were all derived from the polynomials involved in the input. One possible extension would be to consider also the type of relations present and how they are connected logically (likely to be particularly beneficial if problems with more variables or more varied quantifiers are allowed).

A key extension for future work will be the testing of other heuristics. For example the greedy sotd heuristic [23] which chooses an ordering one variable at a time based on the sotd of new projection polynomials or combined heuristics, (where we narrow the selection with one and then breaking the tie with another). We also note that there are other questions of CAD problem formulation besides variable ordering [8] for which machine learning might be of benefit.

Finally, we note that there are other CAD implementations. In addition to QEPCAD there is `ProjectionCAD` [26], `RegularChains` [17] and `SyNRAC` [33] in MAPLE, MATHEMATICA [46] and `Redlog` [24] in REDUCE. Each implementation has its own intricacies and often different underlying theory so it would be interesting to test if machine learning can assist with these as it does with QEPCAD.

## 5 Conclusions

We have investigated the use of machine learning for making the choice of which heuristic to use when selecting a variable ordering for CAD, and quantifier elimination by CAD. The experimental results confirmed our thesis, drawn from personal experience, that no one heuristic is superior for all problems and the correct choice will depend on the problem. Each of the three heuristics tested had a substantial set of problems for which they were superior to the others and so the problem was a suitable application for machine learning.

Using machine learning to select the best CAD heuristic yielded better results than choosing one heuristic at random, or just using any of the individual heuristics in isolation, indicating there is a relation between the simple algebraic features and the best heuristic choice. This could lead to the development of a new individual heuristic in the future.

The experiments involved testing heuristics on 1721 CAD problems, certainly the largest such experiment that the authors are aware of. For comparison, the best known previous study on such heuristics [23] tested with six examples. We observed that Brown's heuristic is the most competitive for our example set, and this is despite it involving less computation than the others. This heuristic was presented during an ISSAC tutorial in 2004 (see Brown [13]), but does not seem to be formally published. It certainly deserves to be better known.

Finally, we note that CAD is certainly not unique amongst computer algebra algorithms in requiring the user to make such a choice of problem formulation. More generally, computer algebra systems (CASs) often have a choice of possible algorithms to use when solving a problem. Since a single formulation or algorithm is rarely the best for the entire problem space, CASs usually use *meta-algorithms* to make such choices, where decisions are based on some numerical parameters [16]. These are often not as well documented as the base algorithms, and may be rather primitive. To the best of our knowledge, the present paper appears to be the first applying machine learning to problem formulation for computer algebra. The positive results should encourage investigation of similar applications in the field of symbolic computation.

## Acknowledgements

## References

1. B. Akbarpour and L. Paulson. MetiTarski: An automatic theorem prover for real-valued special functions. *Journal of Automated Reasoning*, 44(3):175–205, 2010.
2. E. Alpaydin. *Introduction to machine learning*. MIT Press, 2004.
3. D. Arnon, G. Collins, and S. McCallum. Cylindrical algebraic decomposition I: The basic algorithm. *SIAM Journal of Computing*, 13:865–877, 1984.
4. P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.
5. S. Basu. Algorithms in real algebraic geometry: A survey. Available from: `www.math.purdue.edu/∼sbasu/raag_survey2011_final.pdf`, 2011.
6. J. Boyan, D. Freitag, and T. Joachims. A machine learning architecture for optimizing web search engines. In *AAAI Workshop on Internet Based Information Systems*, pages 1–8, 1996.
7. R. Bradford, J. Davenport, M. England, S. McCallum, and D. Wilson. Cylindrical algebraic decompositions for boolean combinations. In *Proc. ISSAC '13*, pages 125–132. ACM, 2013.
8. R. Bradford, J. Davenport, M. England, and D. Wilson. Optimising problem formulations for cylindrical algebraic decomposition. In *Intelligent Computer Mathematics* (LNCS 7961), pages 19–34. Springer Berlin Heidelberg, 2013.
9. J. P. Bridge. Machine learning and automated theorem proving. University of Cambridge Computer Laboratory Technical Report UCAM-CL-TR-792, 2010. Available from: `http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-792.pdf`.
10. J. Bridge, S. Holden, and L. Paulson. Machine learning for first-order theorem proving. *Journal of Automated Reasoning*, pages 1–32, 2014.
11. C. Brown. Improved projection for cylindrical algebraic decomposition. *Journal of Symbolic Computation*, 32(5):447–465, 2001.
12. C. Brown. QEPCAD B: A program for computing with semi-algebraic sets using CADs. *ACM SIGSAM Bulletin*, 37(4):97–108, 2003.
13. C. Brown. Companion to the Tutorial: Cylindrical algebraic decomposition, presented at ISSAC '04. Available from: `www.usna.edu/Users/cs/wcbrown/research/ISSAC04/handout.pdf`, 2004.
14. C. Brown and J. Davenport. The complexity of quantifier elimination and cylindrical algebraic decomposition. In *Proc. ISSAC '07*, pages 54–60. ACM, 2007.
15. C. Brown, M. E. Kahoui, D. Novotni, and A. Weber. Algorithmic methods for investigating equilibria in epidemic modelling. *Journal of Symbolic Computation*, 41:1157–1173, 2006.
16. J. Carette. Understanding expression simplification. In *Proc. ISSAC '04*, pages 72–79. ACM, 2004.
17. C. Chen, M. M. Maza, B. Xia, and L. Yang. Computing cylindrical algebraic decomposition via triangular decomposition. In *Proc. ISSAC '09*, pages 95–102. ACM, 2009.

18. G. Collins. Quantifier elimination for real closed fields by cylindrical algebraic decomposition. In *Proc. 2nd GI Conference on Automata Theory and Formal Languages*, pages 134–183. Springer-Verlag, 1975.

19. G. Collins. Quantifier elimination by cylindrical algebraic decomposition – 20 years of progress. In *Quantifier Elimination and Cylindrical Algebraic Decomposition*, Texts & Monographs in Symbolic Computation, pages 8–23. Springer-Verlag, 1998.

20. G. Collins and H. Hong. Partial cylindrical algebraic decomposition for quantifier elimination. *Journal of Symbolic Computation*, 12:299–328, 1991.

21. N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge University Press, 2000.

22. J. Davenport, R. Bradford, M. England, and D. Wilson. Program verification in the presence of complex numbers, functions with branch cuts etc. In *Proc. SYNASC '12*, pages 83–88. IEEE, 2012.

23. A. Dolzmann, A. Seidl, and T. Sturm. Efficient projection orders for CAD. In *Proc. ISSAC '04*, pages 111–118. ACM, 2004.

24. A. Dolzmann and T. Sturm. REDLOG: Computer algebra meets computer logic. *SIGSAM Bulletin*, 31(2):2–9, 1997.

25. A. Dolzmann, T. Sturm, and V. Weispfenning. Real quantifier elimination in practice. In *Algorithmic Algebra and Number Theory*, pages 221–247. Springer, 1998.

26. M. England. An implementation of CAD in Maple utilising problem formulation, equational constraints and truth-table invariance. University of Bath Department of Computer Science Technical Report 2013-04, 2013. Available from: `http://opus.bath.ac.uk/35636/`,

27. R. Forsyth and R. Rada. *Machine learning: Applications in expert systems and information retrieval.* Halsted Press, 1986.

28. I. Fotiou, P. Parrilo, and M. Morari. Nonlinear parametric optimization using cylindrical algebraic decomposition. In *Decision and Control, 2005 European Control Conference. CDC-ECC '05.*, pages 3735–3740, 2005.

29. H. Hong. An improvement of the projection operator in cylindrical algebraic decomposition. In *Proc. ISSAC '90*, pages 261–264. ACM, 1990.

30. K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

31. Z. Huang and L. Paulson. An application of machine learning to rcf decision procedures. In *Proc. 20th Automated Reasoning Workshop*, 2013.

32. C. Hsu, C. Chang, and C. Lin. A practical guide to support vector classification. 2003

33. H. Iwane, H. Yanami, H. Anai, and K. Yokoyama. An effective implementation of a symbolic-numeric cylindrical algebraic decomposition for quantifier elimination. In *Proc. SNC '09*, pages 55–64, 2009.

34. T. Joachims. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. MIT Press, 1999.

35. T. Joachims. A support vector method for multivariate performance measures. In *Proc. 22nd Intl. Conf. on Machine learning*, pages 377–384. ACM, 2005.

36. D. Jovanovic and L. de Moura. Solving non-linear arithmetic. In *Automated Reasoning: 6th International Joint Conference (IJCAR)* (LNCS 7364), pages 339–354. Springer, 2012.

37. S. McCallum. An improved projection operation for cylindrical algebraic decomposition. In *Quantifier Elimination and Cylindrical Algebraic Decomposition*, Texts & Monographs in Symbolic Computation, pages 242–268. Springer-Verlag, 1998.

38. S. McCallum. On projection in CAD-based quantifier elimination with equational constraint. In *Proc. ISSAC '99*, pages 145–149. ACM, 1999.

39. W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

40. F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

41. B. Schölkopf, K. Tsuda, and J.-P. Vert. *Kernel methods in computational biology*. MIT Press, 2004.

42. F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47, 2002.

43. J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

44. P. Stone and M. Veloso. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8(3):345–383, 2000.

45. A. Strzeboński. Cylindrical algebraic decomposition using validated numerics. *Journal of Symbolic Computation*, 41(9):1021–1038, 2006.

46. A. Strzeboński. Solving polynomial systems over semialgebraic sets represented by cylindrical algebraic formulas. In *Proc. ISSAC '12*, pages 335–342. ACM, 2012.

47. A. Tarski. A decision method for elementary algebra and geometry. In *Quantifier Elimination and Cylindrical Algebraic Decomposition*, Texts and Monographs in Symbolic Computation, pages 24–84. Springer-Verlag, 1998.

48. D. Wilson, R. Bradford, and J. Davenport. A repository for CAD examples. *ACM Communications in Computer Algebra*, 46(3):67–69, 2012.

49. D. Wilson, J. Davenport, M. England, and R. Bradford. A "piano movers" problem reformulated. In *Proc. SYNASC '13*. IEEE, 2013.

50. The benchmarks used in solving nonlinear arithmetic. New York University, 2012. Available from: `http://cs.nyu.edu/~dejan/nonlinear/`