ORIGINAL RESEARCH

Adaptations of Relief for continuous domains of bioinformatics

Andrew Yatsko *

ITMS, the University of Ballarat, Australia

Received: February 10, 2019 **Accepted:** March 27, 2019 **Online Published:** April 10, 2019
DOI: 10.5430/air.v8n1p51 **URL:** https://doi.org/10.5430/air.v8n1p51

ABSTRACT

Relief occupies a niche among feature selection methods for data classification. Filters are faster, wrappers are much slower. Relief is feature-set-aware, same as wrappers. However, it is thought being able to deselect only irrelevant, but not redundant features, same as filters. Iterative Reliefs seek to increase the separation margin between classes in the anisotropic space defined by weighted features. Reliefs for continuous domains are much less developed than for categorical domains. The paper discusses a number of adaptations for continuous spaces with Euclidean or Manhattan metric. The ability of Relief to detect redundant features is demonstrated. A dramatic reduction of the feature-set is achieved in a health diagnostics problem.

Key Words: Feature selection, Relief, Feature weighting, k-NN, Continuous spaces, Data classification, NHANES

1. INTRODUCTION

Filter methods of feature selection assess strength of individual features of the data in predicting the data class.[1–3] As only designated, single-feature sets are evaluated each time, the filter methods are fast, making it possible to quickly exclude any irrelevant features. These features would score very poorly. Although this does not guarantee that the features are irrelevant, filtering them out may be a necessary preprocessing step when it is anticipated that irrelevant features can be many, as in the genome-wide association studies.[3–6] Information gain (IG) from a feature is a popular measure of feature association with data class, a special feature.[1,7,8] Wrapper methods of feature selection assess predictive power of different feature-sets (the subsets of data attributes) with regard to the data class.[1–3] These schemes have a classification method at the core, upon results of which they decide the fitness of a particular feature-set. Engaging a classification method, besides involving many features, is not as simple as the straight application of a class association measure in the filter approach, and therefore wrappers are much slower than filters. Also, because feature combinations are prolific, encompassing them all is impractical. Some concessions are usually made. Particularly, the technique of recursive elimination, starting from the full set of features, can be adopted.[1] One feature is removed at a time based on how adversely it affects the current set fitness. Likewise, one feature a time can be added to an existing feature-set in the snowball rolling fashion, based on how fitness of the resulting set is improved[1] if there exists a sufficiently strong initial set to be reliably evaluated by the classifier at the wrapper core. Wrappers are methods of choice if the aim is to succinctly describe the concept of data. This is because they remove not only irrelevant features but also redundant ones. Same as with irrelevant features, though, the redundancy of discarded features cannot be guaranteed. This cannot be achieved even theoretically as this would mean the concept

*Correspondence: Andrew Yatsko; Email: balunyaan@gmail.com; Address: ITMS, the University of Ballarat, VIC 3353 Australia.

of data is fully known, and so there is no need to learn it. In many cases, the problem of classification can be reduced to the problem of placement of a hyperplane best-separating a pair of data classes in a space swept by numerical features. Working out this hyperplane would involve weighting of features the space is sprung on. This illustrates existence of another class of feature selection methods known as the embedded.[2,3] Generally, the feature-set extracted from a classification method which does its own selection would be the best in a certain sense. Some, but not all classification methods with the embedded feature selection are sensitive to presence of features that would appear mutually redundant, which does not lead to a solution unless a pre-selection is performed or some regularisation applied.[2,9] A good example of embedded feature selection / classification method is the Decision Tree[2,3] although it does not readily weight the features it is eventually spanning. In the standard setting, the decision tree is tolerant to redundancies and reasonably fast to expand.[10] Despite being an embedded method, it actually incorporates the IG filter, although in a non-linear manner. So much so, in the Random Forests framework[2,3,6,11] the classifier merely plays part of the wrapper engine. Despite the search is random, the decision tree use allows for reduction of the search space. This demonstrates different architectures of the feature selection methods and that the modes of processing can be highly intertwined in the quest for higher efficiency.

Feature weighting is a representation aspect of feature selection methods. Usually, higher absolute weights attached to features signify their higher importance, that is, higher informativeness from the view-point of concept learnability. Also, setting some weights to zero effectively excludes corresponding features from the feature-set. The current discourse is particularly interested in the feature weighting within the framework of lazy learners, specifically the Nearest Neighbour classifiers.[12] Although any aspect of data can be binarised, these algorithms are usually applied to data with continuous features as opposed to nominal or discrete. In the space defined by freely changing, numerically represented features, a neighbour to a data-point of choice, that is, instance of data, is another point whose feature values are close to those of the selected instance. The class of the selected point should be similar to that of its neighbour. Of course, the neighbours can be closer to, or further afield from the query instance. Therefore, only the nearest neighbours are drawn to establish the class of the instance concerned. This earned the name k-NN for described methods, where k refers to the size of locally drawn sample. Although k-NN is really a framework, since any classifier can be applied in conjunction with it once the sample is drawn[13] simply counting instances in the sample by class and / or calculating their distances to the instance in focus appeals for its simplicity and can be improved via feature weighting. Usually, the largest class in the sample is assigned to the instance in question, unless there is a tie and then the class altogether least removed from this instance is selected. Higher weights make distances in corresponding directions longer to make value changes count more, as they indeed should for more informative features.

While Relief is a feature weighting method, it actually interprets k-NN to achieve its goals.[4,5,14–17] In Relief, equally sized samples of nearest neighbours are obtained for each class in the two-class setting, and distances are calculated feature-wise to the instance currently in focus whose class is known. The weights are assigned in accordance with the margin each feature is able to exert in separating own class from the opposing class.[18,19] All instances of the data-set are encompassed or a data-wide random selection of test instances is made.[4,15,20,21] Relief is a method of choice in the genome-wide association studies which typically handle the "wide data", where features are encountered in their tens of thousands, but instances hardly score even one thousand[3] due to high acquisition costs. At the same time, only a small cohort of features (genes) is actually responsible for a particular phenotype being studied. It is inconceivable to make use of any of the wrapper methods. It is efficient to run filters to purge a host of irrelevant features. However, the straight answers, if not all in the past, can come only by chance to a lucky few, as the association of many genes rather than a single gene is often responsible for the phenotype in question[3,6] because many genes have the pleiotropic effects. Moreover, a change in a single nucleotide within a gene can be a reason for suppression of other genes, the epistasis. Hence, the actual features in these studies are the single nucleotide polymorphisms (SNP) in the sense of DNA base variation between individuals in a given position of DNA due to the evolutionary factors. Relief occupies a niche between filters and wrappers. It weights features in the context of a given feature-set. So, it will allow to eliminate the irrelevant features. At the same time, it is regarded being insensitive to redundant features[4,15,22] as indeed its progenitor, the k-NN classifier is. This may be a desirable outcome, though, by offering the panoramic view, as opposing to a single-angle view rendered by a wrapper. Relief is affordable in the wide-data setting, as only the calculation of distances between instances is computationally intensive and there are not too many.

Since applications in genomics and proteomics promise significant breakthroughs in the disease treatment or prevention but often have to deal with all-categorical attributes, devel-

opment of Relief went largely in this area.[4] However, the calculation of distances between instances is almost invariably based on the Hamming loss formula, also known as the overlap metric. For a pair of instances, if given a feature the values are different, the partial distance is 1, otherwise 0. The partial results are then summed across the feature-set to obtain the distance. In the genome-wide association context, the feature-wise difference is 0 if the genotypes of two subjects at a SNP are identical, and it is 1 if their genotypes are not identical. The distance between the subjects is then the sum of the differences across all SNPs. More-flexible metrics have been proposed.[6] Despite the simplicity of the Hamming loss, it allows to elegantly circumvent the problem posed by missing values since the difference in a given position of the feature-set is limited. Setting the difference to 1, the maximum, whenever any value in the pair is missing will increase the overall distance and automatically remove instances with many unknowns from the k-NN neighbourhood of a given instance.[8] Albeit, the calculation can be more to the point, having engaged in a guessing game based on value probabilities.[5, 17]

Where data types are mixed, the continuous attributes can be discretised to force all features into the categorical type.[8, 20, 23] The same applies to the classification problems with all-continuous features (explanatory variables) and the regression problems, where additionally the class attribute (response variable) is also continuous.[5] However, the conversion is lossy. Attempts were made to reformulate Relief for all-continuous features, including the class.[17] There are pros and cons of having a problem approached directly in the all-continuous domains. For example, flexibility of the distance function is not an issue as may appear otherwise.[6] At the same time, even if the missing value probabilities could be estimated, a suggested transition from the data space into the space of probabilities is not obvious.[17] Above all, while two unrelated features are incomparable regardless of their type, the value absolute differences for categorical attributes are, but are not for continuous attributes.[24] Rather than summing up the class distance differences to obtain individual feature scores, it was proposed to sum 1-s and 0-s where the differences are positive or negative, respectively, that is, the feature individually classifies or misclassifies the test instance (or vice-versa).[7] However, there may be a loss of sensitivity (in broad terms).

Whether domains are categorical or continuous, there is a problem of setting the threshold below which scores are so low that the concerned features should be discarded. Withholding weak features reduces complexity of the problem, which may be worthwhile despite the features can still be relevant. The solver overall performance, as a trade-off between its computational speed and the admissible error, may increase. Because the scores can be negative, it was proposed that the concerned features are surely irrelevant.[15, 17] However, this does not exclude that there may exist subsets of data where the feature-wise class distance differences are positive. A truly irrelevant feature would yield the score of zero due to value distances being equiprobable by class or between data classes[14] but the opposite is not necessarily true; and then the balance can be shifted one way or the other by chance, since not all of the data is known. Adjacent to the former is the problem of setting the k parameter, as in k-NN, the size of the sample extracted locally for each of data classes[25] since this can affect the calculation of feature scores. Generally, the value of k has to be small, comparing to the amount of data, for the method to be regarded local, that is, taking into account value-combination-specific interactions between features. Although, strictly, a method is local if feature weights vary across the instance space. Increasing k may avert the destabilising influence of possible noise.[17] If k is infinitely increased, though, Relief becomes a global method, with a measure similar to IG, so the method degrades into a filter.[16, 17] However, even with a small k, the impact of instances far away from the perceived boundaries separating classes is much higher than of instances close to the boundaries, because of the incomparable distances. This seems having not received so far a due attention.

The original Relief was formulated for the two-class problems[14, 15] and indeed the majority of diagnostic problems come in this setting. Usually, there are normal subjects, or controls, and subjects affected by a specified abnormal condition. This may have various applications, not only in bioinformatics. In genomics, the population of "wild type" is compared to individuals exhibiting a peculiar phenotype. In the multiclass setting, the problem can be decomposed into two-class subproblems, adopting either the "one against all (others)" or the "pair-wise" classification approach. In the first approach, the class incorporating the rest is sprawling and surrounding the class standing alone, so the decision boundary is complex. In the second approach, the pair-wise decision boundaries are simpler, but the number of problems is quadratic in the number of classes, comparing to the first approach where it is linear.[26] Despite the number of subproblems to solve, the pair-wise approach captures all major twists of the concept and this may be more desirable than handling multiple classes simultaneously. At the same time, perceiving a multiclass problem as a whole, one should be interested in the feature selection applicable to this case. A straight-forward generalisation of Relief, taking the "one against all" view, is to regard all classes different from the test instance own class as the other class.[16] Whether this

or the pair-wise paradigm is embraced, it is assumed that all classes are equally important. If this requirement is relaxed then, for example, bigger classes can be consulted more.[5,17] Be it in two-class or multiclass setting, though, there is a problem of data balancing as the class size may hugely vary, in other words, classes may have much different prior probabilities.[8] In diagnostic problems, unless by design the two classes are equally represented, the controls are usually much more available than the abnormal cases, which spells negative consequences for the results, even though the larger sample is, the better. At least the criterion should be chosen so that any effect of class imbalance is minimised. In Relief this aspect seems to escape scrutiny.

Feature scores calculated by Relief are not the weights in the sense of multipliers in the distance function. Particularly, the scores can be negative. As mentioned, the negative scores are zero-outed. It is possible to normalise the positive scores with their sum and use them as weights.[18,19] Since this leads to improved classification results, the question arises whether it is feasible to recalculate the scores using the weighted distance function in Relief and advance the results even further. Encouraging results were reported by maximising the overall margin, that is, the mean distance difference between the opposite and own class of an instance.[18,19]

2. NOTATION

Let A represent a data-set consisting of M instances on N real-valued attributes, subdivided into C classes, so that (1) holds, where $c = 1 \cdots C$ is the class index. Let a_m be an instance of the data in general, the element of A, $m = 1 \cdots M$.

$$\boldsymbol{A = \bigcup A_c\,;\; M = \sum M_c} \tag{1}$$

The quadratic weighted Euclidean distance between two arbitrary points m_1 and m_2 ($m = 1 \cdots M$) in the instance space is defined by the expression (2) where the point indices '1' and '2' is the shorthand for the mentioned, and $n = 1 \cdots N$ is the attribute index. Without limiting the generality, feature values and so their differences in Eq.2 in position n of the feature-set are normalised by the applicable feature standard deviation s_n.

$$\boldsymbol{d_E^2 = \sum_n \left[w_n \cdot \left(a_{1,n} - a_{2,n}\right)/s_n\right]^2} \tag{2}$$

Likewise, the weighted Manhattan distance is defined by the expression (3). While inessential, s_n can be replaced with the absolute mean deviation in this expression for consistency. The Euclidean and Manhattan distances / space metrics are commonly referred to as L2 and L1 norms, respectively, after the summand exponent in linear expressions for d_E^2 and d_M given by Eqs.2&3.

$$\boldsymbol{d_M = \sum_n w_n \cdot \left|a_{1,n} - a_{2,n}\right|/s_n} \tag{3}$$

The feature weights w_n in Eqs.2&3 are positive, real-valued numbers, subject to the constraint (4), or zero. In Eq.4 Ln is the natural logarithm.

$$\boldsymbol{\sum_{n:w>0} Ln(w_n) = 0} \tag{4}$$

The choice of the Eq.4 constraint is prompted by the notion of k-neighbourhood. In the Euclidean space this neighbourhood is circular in two dimensions if the features are unweighted (have equal weights of 1). If the weights change, the neighbourhood is transformed into elliptical one where the dimension with a larger weight is contracted and with a smaller weight is expanded as the weight inverse. If the sample is small and data is dense, which tends to be more so with more data, the k-neighbourhood in the transformed space will occupy approximately the same area as in the original space. The area is proportional to the product of feature weights. The same holds for the volume in three dimensions and can be proved to hold for any number of dimensions. In the weighted Manhattan space, the k-neighbourhood is diamond-shaped in two dimensions, square if unweighted. Nonetheless, the same considerations apply.

3. ALGORITHM

Initialise all feature weights with 1. Set scores of all features to zero. For each instance a_m find k closets neighbours from each class of the data. Let c_1 point to the class of the current instance and c_2 point to a different class. For each pair of classes indexed c_1 and c_2 find the mean radii r_{1m} and r_{2m} of respective neighbourhoods. Let $r_m = \max(r_{1m}, r_{2m})$ denote the maximum of the two for a pair. Add the quantity (5) in respect of individual features n to their scores.

$$\boldsymbol{[C \cdot (C-1)]^{-1} \cdot \sum_{c \neq 1} \frac{w_n \cdot \left\{\left(\overline{p}_c - \overline{p}_1\right)_{m,n}\right\}_{>0}}{r_m \cdot s_n \cdot M_1}} \tag{5}$$

In Eq.5, the expression under the sum has to be positive to be counted; the overbars denote averaging of 'projections' p across respective class neighbourhoods, where '1' is short for c_1 the class index of a_m, and c is the same as c_2. Each projection contributing to the averages in Eq.5 is a difference of the form $p_{m,n} = |a_{v,n} - a_{m,n}|$ where a_v is an instance from a respective class-dependent k-neighbourhood in the

ISSN 1927-6974 E-ISSN 1927-6982

vicinity of a_m. Find the feature with the largest score. Set weights to zero for features whose scores are less than the highest score by a set factor. Calculate weights of all other features by redistributing evenly any discrepancy in the constraint given by Eq.4, having substituted weights there with the scores. Reiterate a number of times with so obtained initial weights.

Note that in practice more instances than k have to be included in the k-neighbourhood of a specified instance due to the limited precision feature values can be generally obtained with. Particularly, there may be repeating instances. Therefore, after exactly k closest instances are drawn, instances that have the same radius as the farthest instance, but not in the sample, have to be added to the neighbourhood for the statistical evaluation to be fair.

The original Relief[14–17] calculates only the scores which can be used as weights if the negative scores are zero-outed. It does not recalculate the scores. The original Relief score update element, equivalent to Eq.5, is given by the expression (6) using the notation in this article.

$$\frac{(\overline{p}_2 - \overline{p}_1)_{m,n}}{s_n \cdot M} \tag{6}$$

Instead of s_n the maximum-less-minimum for the variable can be used to force all scores into the [-1,1] range. Unlike in Eq.5, no summation is performed in Eq.6 because no provision is made for more than two classes. The multiplier in front of the sum in Eq.5 is a normalisation under the pair-wise classification approach[26] but is not essential. Assuming the two-class setting for the moment, and apart from recalculating the scores and setting weights for the features, the major differences between the proposed method and its archetype are as follows. Firstly, instead of the full margin in Eq.6 for a variable, only the margin gain is accounted for in Eq.5. Any margin loss is not contributing. Put differently, the margin gain is then zero. It can be argued that the more the gain is, the less the loss. Secondly, no locality normalisation, as given by r_m in Eq.5, is applied in Eq.6. This may necessitate setting k too high (of the same order as M).[5,6] Thirdly, the normalisation with regard to the number of instances in Eq.5 is done only in respect of the own class of the test instance, so M in Eq.6 is replaced with M_1 in Eq.5. This is a class balancing element, analogous to reporting sensitivity in respect of the diagnostic condition class and specificity in respect that of controls, instead of the overall success rate (accuracy) regardless of class, when classifying the data.

A depiction of the above algorithm is given in Figure 1. It features three nested loops: for each iteration, for each instance, and ultimately for each feature. Calculations are performed at the beginning, in the middle, or after completion of a loop. Major stop-over points are as shown. This includes setting or revaluation of feature weights at the beginning of each iteration, then feature score initialisation before launching a pass over instances. For each instance class neighbourhoods are then evaluated. The next step is to update feature scores one-by-one. This design would be typical for iterative Reliefs. The input parameters thus include the number of iterations I, of instances M, and of features N. Of course, the input has also to include the data-set itself where M and N are derived from. To compute the class neighbourhoods, the sample size k is required. Additionally, a small number ε in the sensitivity level capacity is required to nullify weights for very small scores, relative to the highest score.

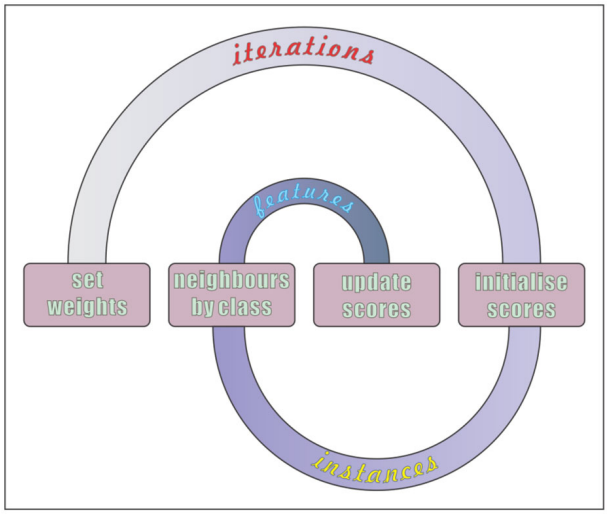

Figure 1. Schematic of iterative Relief cycle

An analogue of the feature score for a feature-set can be obtained by substituting the weighted, normalised projections in Eq.5 with distances. The quantity (7) is the counterpart of Eq.5 in this generalisation, using the notation previously introduced for radii. The total of Eq.7 results over all instances is referred to in this text as the overall, normalised margin gain. This entity concerns the verification agenda and is not directly related to the algorithm. By contrast, the quantity in Eq.5 is the partial, normalised margin gain feature-wise.

$$[C \cdot (C-1)]^{-1} \cdot \sum_{c \neq 1} \frac{\{(r_c - r_1)_m\}_{>0}}{r_m \cdot M_1} \tag{7}$$

When reviewing or benchmarking different algorithms, it is convenient to refer to them by name.[4,5] Main descriptors of the proposed algorithm are that it is iterative, the margin

gain, instead of the whole margin, orientated; it also features a weight constraint interpreted for higher dimensions. Therefore, the algorithm can be named appropriately the 'Imagine' Relief, applying the existing informal convention.

4. Evaluation

Data for around 4,100 participants on 200 attributes was extracted from the US National Health and Nutrition Examination Survey (NHANES) for 2013-14.[27] About half of the features behind the attributes are continuous and the rest are categorical, although mostly binary features. A small number of features are calculated using the core data consisting of demographics, clinical history, anthropometrics, examinations, blood and urine tests, cognitive ability. Statuses were set for the type 2 diabetes mellitus (DM), cardiovascular disease (CVD) and hypertension (HT). The statuses are binary ('yes' or 'no') attributes. The three chronic conditions have vast consequences for health. About 10%-15% of values are missing in the data. These were substituted as previously reported.[28,29] In this evaluation DM exemplifies the class attribute. The prevalence of DM is 18% in the featured population.

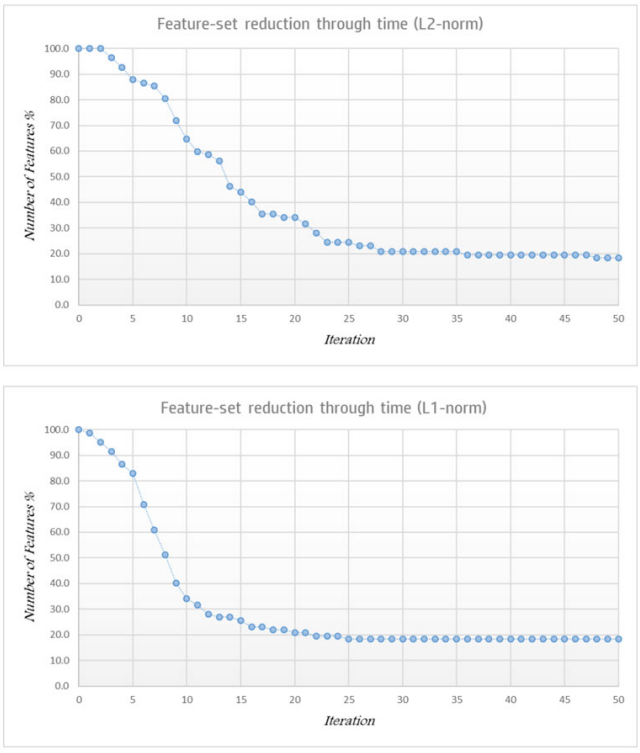

Figure 2. Feature-set relative to initial size at the beginning of algorithmic cycle for L2 and L1 norms

5. Results

The algorithm was run for continuous features only, although excluding a range of known strong and even 'perfect' predictors of DM (all glucose, insulin and tell-tale symptoms related features)[28] to let weaker feature weights evolve. Weights were zero-outed if the scores became 100 times less than the maximum. The sample size k was set to five instances. The results are similar for either distance measure. Over 50 cycles, a feature-set reduction by about 4/5 was achieved (see Figure 2), arriving at the same set regardless of the space metric used. It was observed that the (overall, normalised) margin gain increased, initially fast, and then asymptotically reaching a level about 4.5 times higher than the initial (see Figure 3) be the space metric Euclidean or Manhattan. The feature-set size reduction, although initially slow, exhibited a similar pattern to the margin gain increase (see Figures 2&3).

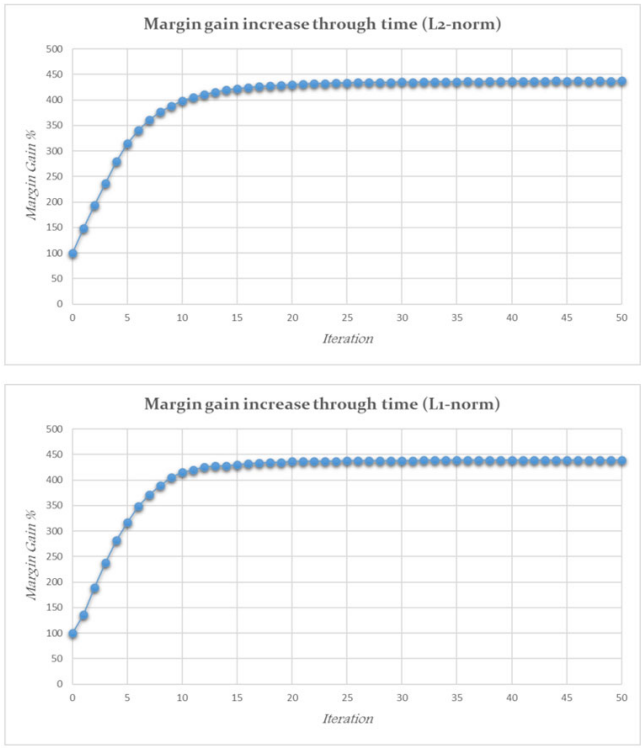

Figure 3. Margin gain at the end of algorithmic cycle for L2 and L1 norms

More features were removed at early stages, and faster for the Manhattan than Euclidean metric, than at later stages, with the feature-set eventually reaching its optimum by size. Table 1 shows the 15 selected features out of the initial 82. Throughout the selection process, weights change slightly if the feature-set does not contract. Some weights increase, while some decrease continuously, eventually becoming very

small. After some feature weights become less than the threshold and nullified, a dramatic change is observed. The major cause of this transformation is the balancing of big weights with small ones in Eq.4. However, despite how this may appear, the feature-set is altogether a different one, and there is a redistribution of weight between features, although consistent with how close the weights were previously. L1-norm weights have a wider range than the L2 counterparts, but if features are arranged by L2 weight as in Table 1, only few minor distortions of order occur in the L1 section, as shown in bold.

A dramatic and fast increase of classification accuracy, comparing to the unweighted feature-set, was observed (see Figure 4). A higher gain was achieved in the Manhattan than Euclidean space. k-NN with a matching distance function to the feature weighting method and k set to five was used for this testing, paired with the leave-one-out validation resampling. However, the accuracy does not change long into the process. The classifier was explained in the introduction. It is applied to the interim results of feature weighting. Zero weight features are effectively switched off. The very slight fall of accuracy past the growth stage can be explained by the gross reduction of the feature-set and a longer time for the weights to adjust in the aftermath.

Figure 4. Classification accuracy by k-NN at the beginning of a specified cycle for L2 and L1 norms

With very little exception, all features in the initial set are thought to be relevant, and no irrelevant features were added on the purpose. Besides, any universally irrelevant features would have been deselected after a single pass. Irrelevant features can confuse a classifier.[13] Conversely, noise can render features irrelevant. Therefore, removal of irrelevant features is expected to lead to more accurate predictions. This is indeed what occurs. Removal of any weakly relevant features should already have negative consequences for the accuracy of classification. Instead, after the weight redistribution, a surplus effect is observed. Even though it gets eventually exhausted, no reversal occurs, while features continue to be removed. It is evident that different equipotent solutions to the problem exist, but the weakest features are not irrelevant ones. In perpetuity, one should expect that all persistent redundant features but their steadfast representatives also get deselected.

Table 1. Feature-set in diminishing L2 weight order at the beginning of 50^{th} cycle for L2 and L1 norms

Feature	L2 weight	L1 weight
Selenium	2.94	5.10
Copper	2.83	3.98
Zinc	2.48	3.39
Cardiovascular risk by BMI	2.23	3.30
Low density lipoprotein cholesterol	2.06	2.71
Atherogenic index of plasma	1.96	2.25
Digit symbol substitution	1.85	2.21
Word recall	1.70	1.73
ApoB lipoprotein	1.40	**1.16**
Cognitive animal fluency	1.29	**1.28**
Sagittal abdomen	1.23	1.13
Word intrusion in recall	0.48	0.16
Waist to height ratio	0.22	0.13
Triglyceride	0.09	**0.07**
Osmolality	0.08	**0.08**

6. Discussion

From the domain knowledge[28,29] the final feature-set in Table 1 contains features regarded to be fair predictors of DM. At the same time, some of their 'duplicates' were removed. For example, there are two versions of the cardiovascular risk but only the stronger one was selected. This version is based on the body mass index (BMI) and prevails because the other one is not corrected for use of medication affecting the cholesterol components replacing BMI in the risk formula.[29] The risk variables are anticipated to top the listing, though, as there is a correction in place for DM. This inconsistency, despite being slight, compromises either version of the risk as a DM predictor variable, but using them in the probe capacity is appropriate here. Also, BMI is a part of the standard protocol for identifying persons predis-

posed to DM.[28] The waist circumference to height ratio (WCHR) is regarded to be a better measure in the same line as BMI.[28, 30] Indeed, WCHR appears in Table 1 and BMI does not. Nevertheless, WCHR is outweighed by the sagittal abdominal diameter, also in Table 1. The diameter is an emergent measure similar to waist circumference.[30] Neither waist circumference nor weight to height ratio, a simplified version of BMI[30] appear in Table 1. So, of the five BMI-like measures only two were retained. Two of the four cognitive function features in Table 1, all preserved so far, come from the same test, but one is auxiliary to the other (word recall / intrusion). In a longer run, features at the bottom of the list get also removed. Evidently, the proposed algorithm has the ability to deselect weaker redundant features in the same line, but it may take indefinitely, if not infinitely, long to have the feature-set reduced to a desired size. Obviously, the end result is also dependent on the initial feature-set.

Relief interprets k-NN, but its fortunes arise from the use of distances rather than the ability to predict classes.[14, 15] Comparing Figures 3 and 4, the highest accuracy of classification is reached just after a handful of iterations. Yet, the capacity to improve on the margin lingers, allowing for the feature selection to continue. With a sufficiently weak set of features, the ability to classify by k-NN should shut completely, yet with Relief it may be still possible to differentiate between features. To a large extent, this applies to the chosen example where the "gold standard" features were intentionally withheld beforehand. This explains the niche status of Relief which, taking away the distance paradox, would have to be placed into the wrapper category.

The original Relief selects the features whose scores are positive, not smaller than a threshold, and the rest is discarded. The features can be weighted according to the scores with weights of discarded ones set to zero. This weighting is supposed to improve the accuracy of k-NN. It seems only natural that, in the weighted distance function context, it is sufficient to reiterate with the weights so obtained[31, 32] to achieve even a better result, especially after normalising them in accordance with Eq.4. This is indeed how the proposed algorithm works. It has been argued that Relief improves the overall margin, as previously explained, and therefore the update procedure should undertake to maximise the margin.[18, 19] The algorithm in the current paper appears to maximise the (overall, normalised) margin gain, and since the accuracy of classification also improves, minimise the loss, although no attempt is made to enforce this. However, applying the scores obtained on each cycle can be seen as moving in the direction of gradient of the margin gain, as the objective of optimisation; while the redistribution of any discrepancy in Eq.4 after substituting the feature scores for weights interpreted as the rate-of-advancement setting arrangement on the path to the maximum.

An alternative listing is presented in Table 2. These are 15 (out of 82) features whose IG was the highest[8] to match the number of features in Table 1. The IG results were converted to weights using Eq.4. In both tables, three features that do not appear in the other table are dimmed out. As expected[8] IG approximates fairly the reduced feature-set, both component- and arrangement-wise, but without limiting redundancy. For example, both cardiovascular risk variables are listed; the waist circumference, its ratio to height, and the sagittal abdominal diameter are all listed. Of note, the very action of filtering out reduces redundancy of the feature-set because some of the redundant features are also 'weakly' relevant as are, for example, BMI and weight to height ratio that do not show up in Table 2.

Table 2. Top 15 features weighted by information gain

Feature	Weight
Cardiovascular risk by BMI	2.34
Cardiovascular risk by cholesterol	2.05
Selenium	1.55
Atherogenic index of plasma	1.31
Triglyceride	1.30
Low density lipoprotein cholesterol	1.09
Waist to height ratio	0.94
Sagittal Abdomen	0.90
Copper	0.85
Waist Circumference	0.81
Digit symbol substitution	0.77
Osmolality	0.71
Cognitive animal fluency	0.69
Zinc	0.59
Urinary albumin to creatinine ratio	0.56

Weighted by IG, the cardiovascular risk variables are topping the list in Table 2. It seems phenomenal that only one of them is retained when weighted by the Imagine Relief, as evident from Table 1. Ousting of the second variable, which by itself is only slightly weaker a version the first one, occurs in a number of steps with its weight gradually reduced until it becomes infinitesimal and nullified. The iterative component of the proposed algorithm is thus essential for the feature-set redundancy reduction. In this connection, it may seem counterintuitive that all cognitive function variables coalesce in Table 1, including the mentioned two from the same test that do not appear in Table 2. However, same as the metals in Table 1, the three cognitive tests targeting, in the order of listing, the abilities to concentrate, to memorise, or to bring to mind - are all functionally different.

The k-NN accuracy of the weighted feature-set from Table 2

is inferior to that from Table 1. The comparison is given in Table 3.

Table 3. k-NN accuracy for different feature selection methods and space metrics

Method	Space	Sensitivity %	Specificity %
IG	Euclidean	69.4	96.5
Imagine	Euclidean	73.5	96.9
IG	Manhattan	75.0	97.3
Imagine	Manhattan	78.8	98.1

As previously noted, the multiclass setting is not typical for the diagnostic problems. The data example brought up in this paper is no exception despite its versatility. Hence, the provision in Eq.5 for situations when there are more classes than two is yet to be exercised. However, the proposed treatment which follows the path of pair-wise classification[26] is different from the approaches pursued elsewhere.[5, 16, 17]

Relief algorithms can be fooled by many features being weak or irrelevant, which may affect the calculation of distances.[15] This is offset by drawing larger samples of nearest neighbours by class, an antinoise measure.[17] Also, noise can be separately treated.[13] Same as other Reliefs, the proposed method thrives on features having contrasting strengths.[15]

The data has to be in full supply for the algorithm to work. Methods of data completion in situations when some of it is missing were previously developed[8, 28, 29] and some indeed applied to the test data in the current work.

REFERENCES

[1] Mlambo N, Cheruiyot WK, Kimwele MW. A survey and comparative study of filter and wrapper feature selection techniques. International Journal of Engineering and Science (IJES). 2016; 5(8): 57-67.

[2] Jovic A, Brkic K, Bogunovic N. A review of feature selection methods with applications. Proceedings of the 38-th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2015; 1200-5. IEEE. https://doi.org/10.1109/MIPRO.2015.7160458

[3] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23(19): 2507-17. PMid:17720704. https://doi.org/10.1093/bioinformatics/btm344

[4] Urbanowicz RJ, Meeker M, La Cava W, et al. Relief-based feature selection: introduction and review. Journal of Biomedical Informatics. 2018; 85: 189-203. PMid:30031057. https://doi.org/10.1016/j.jbi.2018.07.014

[5] Urbanowicz RJ, Olson RS, Schmitt P, et al. Benchmarking Relief-based feature selection methods for bioinformatics data mining. Journal of Biomedical Informatics. 2018; 85: 168-188. PMid:30030120. https://doi.org/10.1016/j.jbi.2018.07.015

[6] Arabnejad M, Dawkins BA, Bush WS, et al. Transition-transversion encoding and genetic relationship metric in ReliefF feature selection improves pathway enrichment in GWAS. BioData Mining. 2018; 11(23). https://doi.org/10.1186/s13040-018-0186-4

[7] Bagirov A, Yatsko A, Stranieri A, et al. Feature selection using misclassification counts. Proceedings of the Ninth Australasian Data Mining Conference (AusDM). In: Conferences in Research and Practice in Information Technology (CRPIT). 2011; 121: 51-62. ACS.

[8] Jelinek HF, Yatsko A, Stranieri A, et al. Diagnostic with incomplete nominal/discrete data. Artificial Intelligence Research. 2015; 4(1): 22-35. https://doi.org/10.5430/air.v4n1p22

[9] Zou H, Hastie T. Model building and feature selection with genomic data. Liu H, Motoda H (editors) Computational Methods of Feature Selection 2008; 393-411. Chapman & Hall / CRC.

[10] Quinlan R. C4.5: programs for machine learning. 1993. Morgan Kaufmann.

[11] Breiman L. Random forests. Machine Learning. 2001; 45(1): 5-32. https://doi.org/10.1023/A:1010933404324

[12] Wettschereck D, Aha DW, Mohri T. A Review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review. 1997; 11: 273-314. https://doi.org/10.1023/A:1006593614256

[13] Segata N, Blanzieri E, Delany SJ, et al. Noise reduction for instance-based learning with a local maximal margin approach. Journal of Intelligent Information Systems. 2010; 35(2): 301-31. https://doi.org/10.1007/s10844-009-0101-z

[14] Kira K, Rendell LA. The feature selection problem: traditional methods and a new algorithm. Proceedings of the Tenth National Conference on Artificial Intelligence. 1992; 2: 129-34. AAAI.

[15] Kira K, Rendell LA. A practical approach to feature selection. Proceedings of the Ninth International Conference on Machine Learning. 1992; 249-56. Morgan Kaufmann.

[16] Kononenko I. Estimating attributes: analysis and extensions of Relief. Proceedings of the European Conference on Machine Learning. 1994; 171-82. Springer.

[17] Kononenko I, Kukar M. Machine learning and data mining: introduction to principles and algorithms. 2007. Horwood.

[18] Sun Y. Feature weighting through local learning. In: Liu H, Motoda H (editors) Computational Methods of Feature Selection 2008; 233-53. Chapman & Hall/CRC.

[19] Gilad-Bachrach R, Navotz A, Tishby N. Margin based feature selection - theory and algorithms. Proceedings of the 21-st International Conference on Machine Learning. 2004; 43-50. ACM.

[20] Blessie EC, Karthikeyan E. ReliefDisc: an extended relief algorithm using discretization approach for continuous features. Proceedings of the Second International Conference on Emerging Applications of Information Technology (EAIT). 2011; 161-4. IEEE. https://doi.org/10.1109/EAIT.2011.39

[21] Dash M, Yee OC. ExtraRelief: improving Relief by efficient selection of instances. Proceedings of the Australasian Joint Conference on Artificial Intelligence. In: Orgun MA, Thornton J (editors) AI 2007: Advances in Artificial Intelligence. Lecture Notes in Computer Science. 2007; 4830: 305-14. https://doi.org/10.1007/978-3-540-76928-6_32

[22] Yang J, Li YP. Orthogonal Relief algorithm for feature selection. Proceedings of the International Conference on Intelligent Computing. 2006; 227-34. Springer.

[23] Demsar J. Algorithms for subsetting attribute values with Relief. Machine Learning. 2010; 78(3): 421-8. https://doi.org/10.1007/s10994-009-5164-0

[24] Chikhi S, Benhammada S. ReliefMSS: a variation on a feature ranking ReliefF algorithm. International Journal of Business Intelligence and Data Mining. 2009; 4(3-4): 375-90.

[25] McKinney BA, White BC, Grill DE, et al. ReliefSeq: a gene-wise adaptive-k nearest-neighbor feature selection tool for finding gene-gene interactions and main effects in mRNA-Seq gene expression data. PloS One. 2013; 8(12): e81527. https://doi.org/10.1371/journal.pone.0081527

[26] Park SH, Furnkranz J. Efficient pairwise classification. Proceedings of the Eighteenth European Conference on Machine Learning (ECML). 2007; 658-65. Springer-Verlag.

[27] National health and nutrition examination surveys. Available from: http://cdc.gov/nchs/nhanes/

[28] Stranieri A, Yatsko A, Jelinek HF, et al. Data-analytically derived flexible HbA1c thresholds for type 2 diabetes mellitus diagnostic. Artificial Intelligence Research. 2016; 5(1): 111-34. https://doi.org/10.5430/air.v5n1p111

[29] Venkatraman S, Yatsko A, Stranieri A, et al. Missing data imputation for individualised CVD diagnostic and treatment. Computing in Cardiology. 2016; 43: 349-52. IEEE.

[30] Yatsko A. Indexing adult obesity by waist-to-height and weight-to-height ratios. Journal of Biomedical Engineering and Informatics. 2017; 3(2): 20-35. https://doi.org/10.5430/jbei.v3n2p20

[31] Moore JH, White BC. Tuning ReliefF for genome-wide genetic analysis. Proceedings of the European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. 2007; 166-75. Springer.

[32] Draper B, Kaito C, Bins J. Iterative Relief. Proceedings of the Computer Vision and Pattern Recognition Workshop (CVPRW). 2003; 6: 62. IEEE. https://doi.org/10.1109/CVPRW.2003.10065