ORIGINAL RESEARCH

Weighting features by the value displacement rebound

Andrew Yatsko*

ITMS, the University of Ballarat, VIC, Australia

Received: March 25, 2020 **Accepted:** July 6, 2020 **Online Published:** July 29, 2020
DOI: 10.5430/air.v9n1p27 **URL:** https://doi.org/10.5430/air.v9n1p27

Abstract

Learning from examples draws on similarity, a concept which formalisation leads to the notion of instance space. Continuous spaces are easier to embrace since, unlike discrete, they often can be seen as hyper-constructs of 3D. Unsurprisingly, the instance-based learning methods are more developed for continuous domains than for discrete ones. The value difference metric (VDM) is one of the few examples of metrics for discrete spaces. Mixed reports about utility of VDM exist. In this paper VDM is compared with another approach where data features are weighted by the Information Gain. Some vulnerabilities of VDM are identified. A weighting method, nothing like VDM, although inspired by the former, is proposed. The results are in favour of the new weighting scheme with illustration of utility for health diagnostics.

Key Words: Feature selection, Feature weighting, Class imbalance, Discrete spaces, VDM, Diagnostics, NHANES

1. Introduction

Stanfill and Waltz[1] have introduced a metric for spaces spanning symbolic / discrete features that accounts for differences in value class frequencies by feature, given two instances of data. They named it the Value Difference Metric (VDM). Later, Cost and Salzberg[2] have modified the metric by removing a term which rendered it asymmetric and by penalising distances to some instances to reduce effects of noise. The second requires another pass through the data and can be considered a general method outside the competence of space metric.[3,4] Cost and Salzberg noted that their approach lends itself easily to parallel computations, therefore naming it the Parallel Exemplar-Based Learning System (PEBLS).[2] They however acknowledged that this may appear overreaching and would equally apply to the original method of Stanfill and Waltz.[1] As far as VDM itself is concerned, the modification[2] rendered it simpler. In this form (setting a new standard[5,6]) the feature-wise distances are calculated by Eq.1.

$$\delta(F;f_1,f_2) = \frac{1}{2}\cdot\sum_{t=1}^{T}|P_1 - P_2|$$

$$P_1 = P(C = c_t|F = f_1)$$

$$P_2 = P(C = c_t|F = f_2) \qquad (1)$$

In Eq.1, C and c_t are the class and its particular value; likewise F is a feature and f_1 and f_2 are its two values being compared; $T > 1$ is the number of classes (object types) and t is the class index. While it is called the value difference metric, in fact, the relative class frequencies / probabilities P, given one value and then the other, are subtracted. The prior squaring of the summand in Eq.1[1] was also abolished in this version.[2,5]

This approach exploits redundancies which exist on the value level within a feature. The feature redundancy is a topic often discussed, but each value of a symbolic feature can be cast

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

as a binary feature of its own. Outwardly, this may seem strange as one can expect uniqueness of values. However, in practice the likeness of categories is wide-spread. Consider the classification of race / ethnicity in USA demographics. There are, for example, the Mexican and the Hispanic but not Mexican categories. Due to the genetic likeness, the two ethnicities would exhibit similar patterns of succumbing to a disease. So, the same proportion of Mexicans and other Hispanics is expected to be affected, or not, by a particular health condition. This is despite the frequencies of the two population categories are different. Both Latinos and blacks (African descent) are at the risk of diabetes but not whites (European descent).[7] Compare now the black and white population cohorts. Unlike before, there are will be differences in relative frequencies of diabetics and non-diabetics between the two. Albeit, this leaves no room for intermarriages. Anecdotally, the marital status also poses a dilemma: "married" would be similar to "partnered" and "divorced" to "separated", whilst "married / partnered" would be dissimilar to "divorced / separated", regardless how the data is dissected, so long this is related to the goal (singles would have a higher propensity to fast food than functional families in the case of diabetes). The value differences will be smaller in the first instance and larger in the second for this feature.

Generally, it is unknown in advance that some values can be similar and other dissimilar in their midst. While Stanfill and Waltz[1] took their inspiration from the task of converting graphemes to phonemes where the confusables like 'c'&'k', 'e'&'i', 'g'&'j', 'i'&'y' are abundant, the same applies to amino-acids in a protein chain in predicting its space folding formation-a challenge that Cost and Saltzberg[2] have taken upon assuming no prior knowledge. Neither the occurrence of similar values is paramount. A contrast between certain values is intrinsic in classification problems. To illustrate, say, what do 'red' and 'green' have in common - both describe 'apple' - but not so much does 'blue'.[8] Eq.1 takes care of this all by examining the relative frequencies. Also, multiple features are usually involved. Therefore, the complete metric adds up the individual feature effects as in Eq.2.

$$d(x_1,x_2) = \frac{1}{N}\cdot\left(\sum_{n=1}^{N}\delta_n{}^{s}\right)^{1/s} \tag{2}$$

In Eq.2 x_1 and x_2 are two feature vectors being compared, each having N attributes indexed n; s is an exponent usually taking values 1 or 2 hence giving the expression likeness of the Manhattan or Euclidean distance, respectively.[8] (For clarity, component-wise x_1 is $(x_{1,1}...x_{1,n}...x_{1,N})$ and so is x_2 by analogy; δ_n is obtained by substituting $x_{1,n}$ and $x_{2,n}$ for feature F_n values in Eq.1.) Smaller individual distances δ lead to a smaller overall distance d between the two tuples of feature values in vectors x_1 and x_2 also referred to as points in the instance space. The smaller individual distances are, the greater is the overall similarity between the two involved instances of data.

Despite plausibility of the above, mixed views about utility of the scheme were held in the past. One downside is that computationally it is challenging.[2] Otherwise, if some features are continuous they can be discretised[9] or the specific value probabilities can be interpolated from adjacent ranges[8] so this is hardly a limitation. If nothing else, VDM amongst its peers offers a better deal for discretised ranges due to its proximity sensing ability. Yet, comparing to other methods, no independent observer seemed to claim a clear advantage of VDM.[10, 11] Nonetheless, more recently, VDM found its application in one instance-based learning approach.[6]

The formulation in Eq.1&2 is consistent across a number of publications[5, 6, 8, 9] with a minor variations as to normalisation of Eq.2 inconsequential to the end result. It is being identified as either VDM, PEBLS, MVDM, or SVDM. SVDM[5, 6] (simplified VDM) is the closest match because PEBLS and MVDM (modified VDM)[2] both imply additional weighting of instances and PEBLS also using parallel computations, but none is being reported. It seems more appropriate to use VDM in the sense of evolving concept, though, because neither 'modified' nor 'simplified' reveal the nature of change applied to the scheme.

VDM, coupled with the Nearest Neighbour paradigm, creates a classifier whereby an unlabelled instance is assigned to the class more represented in a vicinity of the instance as a point in the space. The vicinity radius is determined by the number of closest instances drawn k which can be as small as 1. Therefore, the method goes by the name of k-NN. If there is a tie, the class closest overall to the unlabelled instance is usually chosen.[12] Some approaches discount contributions of more distant nearest neighbours by using various $kernels$[12] particularly the distance inverse.[10, 11]

A variety of metrics can be plugged into k-NN, particularly metrics using the feature weighting, with the feature-wise distances defined in Eq.3.[10, 11]

$$\delta(F;f_1,f_2) = \begin{cases} w_F, f_1 \neq f_2 \\ 0, f_1 = f_2 \end{cases} \tag{3}$$

In Eq.3 w is the weight (a real non-negative constant) assigned to feature F. If uniformly $w = 1$ for all features and $s = 1$ the calculation in Eq.2 is known by the name of Hamming loss / distance[12] also referred to as the overlap metric.[10] The partial distance is w if the two values of a given feature are different, and zero if they match.

In this work VDM is compared with the feature weighting by the Information Gain (IG).[4, 10] If the weights are normalised by IG of the class variable then due to dividing by N the result of Eq.2 cannot be more than 1[12] which is computationally as well as theoretically convenient. The same is true of VDM due to division by 2 in Eq.1 introduced here (it can be verified the sum there is no more than 2).

The expression for IG weights is given in Eq.4.

$$w_F^{IG} = 1 + \frac{H(F)}{H(C)} - \frac{H(C \times F)}{H(C)}$$
$$H(F) = -\sum_{v=1}^{V_F} \left[P_{F,v} \cdot log_2 P_{F,v}\right]$$
$$P_{F,v} = P(F = f_v) \tag{4}$$

In Eq.4 H is the entropy – a quantity representing the average information contained in a feature; it is expressed in terms of value probabilities P. Three features are involved: the class variable C, an explanatory feature F with regards to C, and $C \times F$ – the Cartesian product of C and F. The latter is an artificial feature whose values are unique combinations of values of the former two, which is implicit of using joint value probabilities for C and F in calculation. Only $H(F)$ is shown, where v is the value f index running up to the number of values V for feature F; $H(C)$ and $H(C \times F)$ are expressed by analogy.

Features that are more relevant to the problem should have higher weights[10, 11] making any shifts in their directions more sensitive (tending to cross class boundaries). Vice-versa, shifts in directions of less relevant features should be desensitised by assigning smaller feature weights (so as to prevent crossing the boundaries by chance). Thus, altogether irrelevant features should have the weight of zero. Obtaining weight one way or another can be used to filter out irrelevant features.

VDM seemingly adjusts to feature strength despite not being a weighting method. Indeed, if a feature is irrelevant then the class probabilities should be no different from value to value, therefore reducing the partial distance δ given by Eq.1 to zero.[5] Instead, for a perfect predictor – a feature as relevant as the class variable – it can be verified that δ is reaching its maximum of one (unity) whenever the values being compared belong to different classes, similar to IG. If a predictor is not perfect, some frequency leakage will manifest, giving rise to a range of values between 0 and 1 for δ. If a predictor is weak, a high value mix can be expected. This spells closer class probabilities for two arbitrary values of the variable, so the result of Eq.1 would approach zero. In view of the above it is unsurprising that VDM weighted variants, as was originally held[1] but, as mentioned, considered problematic and made redundant,[2] are hard to come by.[10]

Next, an alternative to IG weighting is introduced, inspired by VDM.

2. Method

VDM attracts for its perspicuity and flexibility (Eq.1).[2] Handling of probabilities in IG is more intricate (Eq.4) and the weighting approach is intrinsically rigid (Eq.3). Unlike in *local* weighting, which applies to VDM to some extent, the *global* weights do not change with instance.[10] At the same time, there were reports that weighting by IG could outperform VDM.[10] Claim verification is in order and also this invites rethinking of VDM. In this wake, a weighting method is proposed where the notion of value 'differences' is seen in a different light.

Consider a two-class situation (binary classification). A feature value could perfectly predict a class if it was entirely within boundaries of that class (feature values can be cast as numeric binaries). Therefore, the smaller of the relative value frequencies between the two classes is the "value displacement". The other part is the 'rebound'. The total of rebounds over all values - the Value Displacement Rebound (VDR) - weighs the feature relevance in accordance with Eq.5.

$$w_F^{VDR} = 1 - \sum_{v=1}^{V_F} \min\left[P_1, P_2\right]$$
$$P_1 = P(F = f_v | C = c_1)$$
$$P_2 = P(F = f_v | C = c_2) \tag{5}$$

Eq.5 uses the same notation as previously. Particularly, $P(\cdot|\cdot)$ represents conditional probabilities. When $T > 2$ (multiclass setting) the feature weight can be generalised as in Eq.6.

$$\overline{w}_F = \frac{2}{T \cdot (T-1)} \cdot \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} w_F(c_i, c_j) \tag{6}$$

That is, by encompassing all two-class problems and taking the mean. While Eq.6 is intended here for VDR, a variety of feature weighting schemes could adopt the same approach. However, IG weights given by Eq.4 are self-sufficient.

Other than Eq.6 and focusing on relevant attribute selection (feature ranking, dimensionality reduction) rather than weight application, this method is equivalent to a little known technique by Baim.[13]

3. Evaluation

Data for around 6,850 participants on 290 attributes was extracted from the USA National Health and Nutrition Examination Surveys (NHANES) for years 2011-14.[14] About 40% of the features behind the attributes are continuous and the rest are categorical, although mainly binary features. The core content is demographics, clinical history, anthropometrics, examinations, blood and urine tests, cognitive ability, etc. A small number of features are aggregates using the core data. About 25% of values are missing in the data. These were substituted as previously reported.[15] For the sake of evaluation, continuous attributes were discretised by the even (equal) frequency method advanced previously.[4] The number of intervals was set to five for all real-valued variables[16, 17] unless the algorithm had to reduce it. Generally, this number has to be small for reliable estimation of probabilities but not too small so as not to lose the discriminative power. Otherwise, if a variable allowed dual representation, its discrete form was used.

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 the health. The prevalence of DM is roughly 20%, CVD 40% and HT 45% in the featured population who are 35+ year-olds. The aforementioned statuses have the designation of class variables in three diagnostic tasks that were attempted. In classifying by k-NN the parameter k was set to 3 regardless of implementation or feature listing (exclusive of equidistant neighbours at the vicinity fringe also drawn). No space warping was considered, so the parameters in Eq.2 was invariably 1.

Three different feature-set sizes were attempted. The full feature-set for a problem excludes only a small number features that one way or another are involved in setting of the corresponding class variable. The essential is a subset of the full feature-set that additionally excludes DM, CVD and HT statuses, unless they are the class variable, and all variables expressly linked to them. The three statuses are closely interrelated[15] therefore launching a circular argument if one is to be determined using others, not all of them known. For example, the DM Status is instrumental in computation of a certain CVD Risk, so both are excluded from the feature-set when predicting the occurrence of CVD. Also, the essential feature-sets do not include what can be regarded “insider knowledge” such as frequency of doctor visits, indications for treatment, etc. Additionally, the essential feature-sets exclude 10% of the database all features poorly scoring by IG in respective problems, and some known redundancies.

The short feature-set is obtained by selection and is a subset of the essential one. Fifty features were selected to classify whether DM or CVD, and forty to classify HT; that is, under 20% of the original set.

4. Results

Table 1 shows the accuracy of DM, CVD or HT prediction when classifying by k-NN with VDM, or the overlap metric weighted either by IG or the proposed method ‘Rebound’ (VDR). Reported are the accuracy aspects sensitivity and specificity, that is, the success rate in predicting a chronic condition or its absence, respectively. The balanced accuracy, which is their mean, is graphically represented in Figure 1.

Table 1. Aspects of k-NN accuracy for different metrics and selections of features

Method / Feature-set *		Specificity (%)			Sensitivity (%)		
		DM	CVD	HT	DM	CVD	HT
VDM	F	98.6	96.0	95.5	74.4	86.8	86.1
VDM	E	98.2	94.5	93.8	72.8	81.9	83.9
VDM	S	97.4	92.2	90.5	77.7	81.6	83.7
IG	F	97.2	94.1	94.9	84.2	88.1	90.6
IG	E	96.1	92.7	91.9	80.7	87.2	87.7
IG	S	96.0	90.9	89.4	79.9	84.5	86.9
Rebound	F	98.0	96.5	96.9	87.9	93.4	93.1
Rebound	E	97.0	94.7	94.4	84.6	90.5	91.9
Rebound	S	97.4	92.8	92.0	86.5	89.9	89.7

‘F’-Full, ‘E’-Essential, ‘S’-Short

The accuracy is expected to drop when replacing the full with essential feature-sets because a number of powerful predictors engaging the other statuses become unavailable. This was indeed observed in all problems.

The objective in feature selection was to prop the minimum of sensitivity and specificity instead of the balanced accuracy. Yet, while a gain of accuracy is desirable, shortening of the feature-set is higher on the agenda. Also, the feature selection method has its own bias, and the problems in consideration are all different. With this recourse, after short-listing the features, the gain of accuracy realised for DM but did not for CVD or HT (see Figure 1). Nonetheless, if the accuracy is gained, as observed, this is more on the sensitivity rather than specificity side; and the other way around if lost, commensurate with the set size (see Table 1). To represent this more clearly, suppose a_0 is specificity and a_1 sensitivity; then the sensitivity ‘lag’ can be calculated as follows $(a_0 - a_1)/(a_0 + a_1)$. The effect is depicted in Figure

2. The lag is uncontained only for IG in the DM and CVD problems.

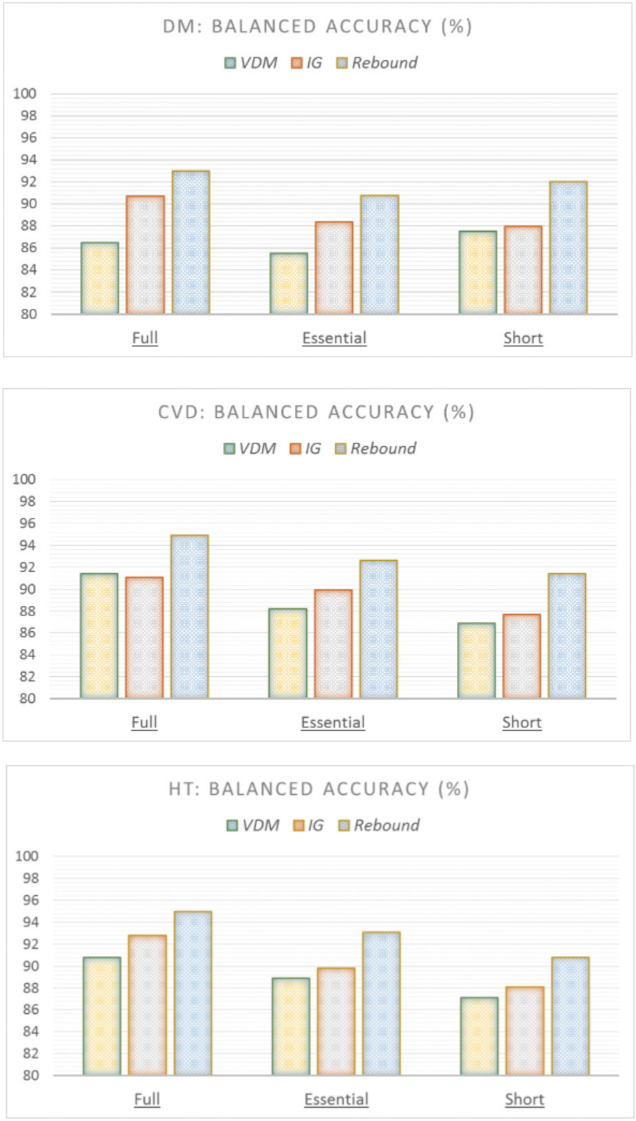

Figure 1. Balanced Accuracy by VDM, IG or VDR (Rebound) for differently sized feature-sets

None of the methods slipped under the 80% balanced accuracy mark, which is notionally 'great'.[15] The accuracy estimate is obtained by leave-one-out cross-validation, that is, by testing all instances in turn against rest of the data.[12] This is a standard approach but may appear optimistic. Overall, paired with k-NN, VDM is less assuring when compared to the weighting by IG, and the latter underperforms comparing to VDR.

The *class imbalance* (also 'skew')[3, 4, 18–20] is more profound in the DM problem than the two others. The prevalence of DM is roughly half of that for CVD or HT. This undoubtedly gnaws away at the sensitivity of DM prediction. For VDM the result is never above 80%, which is also true for IG on the short feature-set (see Table 1). Overall, the gap between sensitivity and specificity is wider in DM than in CVD or HT, scaling almost linearly with class imbalance (see Figure 2). Again, by this measure, VDM appears coping less with class imbalance than IG, and IG less than VDR (see Figure 2).

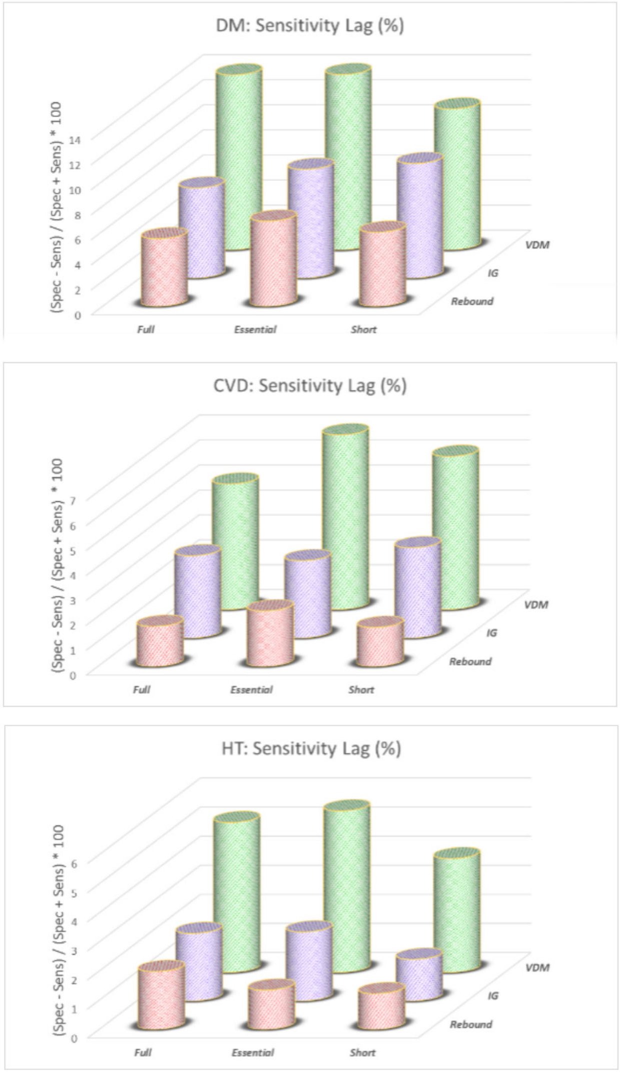

Figure 2. Sensitivity Lag by VDM, IG or VDR (Rebound) for differently sized feature-sets

5. Discussion

5.1 Shortlisting the features

The short feature-sets were obtained from the essential ones by the sequential backward elimination[5] wrapping around the Naive Bayes (NB) classifier.[16,17] The sizes were selected so that the accuracy of NB peaks and does not favour sensitivity nor specificity. NB is a classifier that is naturally 'uncomfortable' with redundant features.[4,12,16,17] However, it can carry on with irrelevant ones. This does not impede the *wrapper* reckoning, but to disambiguate and speed up, the weakest features were removed from the essential feature-sets. Also, for more informed decisions, the performance of individual features in previous deletion cycles was being taken into account. A *wrapper* is generally thought being able to rid of irrelevant as well as redundant features.[21–24] Irrelevant features can be confused with the weak, potentially misdirecting the search.[21,22] The ability of NB to sense redundancies is additional to that of a wrapper. Since not all feature combinations get encompassed under the sequential backward elimination, far from so, the heuristic search (*breadth first hill climbing*[12]) thus becomes more to the point. Another reason to choose NB is that it is fast and can be made even faster if not to recompute the involved probabilities. The probabilities in NB are obtained from all data, therefore rendering single instance contributions infinitesimal under the leave-one-out cross-validation. In effect, the classifier accuracy to guide the selection of features is obtained by *resubstitution*. However, instead of recomputing, a provision can be made to only amend probabilities as necessary.[16]

When selecting features there is a trade-off between their number and the accuracy. If features are gradually deselected, as in the described NB wrapper, the accuracy may initially increase as data noise and computation volume are reduced. The irrelevant features distract the classifier at training. Besides, lengthy calculations are less precise. So, the accuracy does not necessarily fall from the start. Thereafter it degrades, nonetheless, but usually slowly with many features. This is the general trend, but there are ups and downs throughout as features interact slightly differently in reduced sets. Therefore, only the immediate goal can be pursued (hill climbing). Also, there is a tendency to increasing gap between sensitivity a specificity. The sensitivity wears off faster. This is on the part of class imbalance.[3,4,18–20] So, the objective presently was to prop the minimum of the two instead of the balanced accuracy.

5.2 VDM vulnerabilities

A demise of VDM sensitivity in predicting DM comes to the fore (see Table 1). DM is more class-imbalanced than CVD or HT. One can observe that while any classification problem can be reduced to a series of binaries, variables, especially discretised continuous ones, usually have more values than two. Therefore, there can be small value frequencies, more so by class. The class imbalance clearly exacerbates the situation since a predictor is good if its values are distinct. Small frequencies make estimation of probabilities less reliable. Elsewhere, the sensitivity of VDM to class imbalance was also noted.[8]

It had been proposed that VDM relies on a conditional independence of features, same as NB, although perhaps differently defined.[17] Since the k-NN accuracy has improved after short-listing the features in DM, this does not thwart the argument but is inconclusive. In two other problems the accuracy by VDM only fell. Overall, the effect does not manifest strongly; although, the feature selection relies on NB, and to stage this properly requires k-NN with VDM. Also, the latter has a lead over the former to the tune of 10% in the CVD and HT classifications and 5% in DM (behind the scenes). It seems, VDM cannot complain about the accuracy on the same grounds as NB. Interestingly, NB fared overall some 5% better on the DM task than on CVD or HT. This is despite having a term quantifying the class imbalance in its formulation. For NB, the violation of feature independence, given a class, condition is indeed the defining moment in its lack of success.[16,17]

To dissuade the premise of VDM reliance on a conditional feature independence more, k-NN uses a small locally extracted sample, unlike NB that draws on all data. The probabilities in Eq.1 are also globally extracted, but this is not the same as the distances that underlie the true probabilities involved. Generally, the nearest neighbour algorithms are known to tolerate redundancies.[11] Particularly, training redundant feature weights by various methods of performance feedback was demonstrated to increase the accuracy of predictions.[10] This suggests an independence existing on the local level. Indeed, because of the 'frozen' state all sources of influence of one variables over others cease.

5.3 Class imbalance

The class imbalance is not so much of a concern for machine learning as it is for data mining and knowledge discovery. In machine learning the focus is on algorithms, their optimal set of parameters and how they perform under stress.[12] Particularly of interest is the learning curve for small samples or how the algorithms computationally scale with data. One aspect of this is the class imbalance, but realistically only the application side suffers. In data mining and knowledge discovery the focus is on describing the *data concept*.[12] Class imbalance is notionally related to *class noise*.[3,4] It is not possible to learn concepts reliably in noisy environments. Perhaps it is possible to reformulate the problem in study[3] but this implies the *domain* knowledge.

The theoretical impasse the class imbalance presents is that at certain its levels it is impossible to beat the majority rule by fine-tuning the algorithm or replacing it altogether. This is because, unless the concept is fully known, neither the *classifier bias* nor the *data variance* can be reduced to zero. Class imbalance at moderate levels can be remedied by undersampling of the majority class, oversampling of the minority class, or both.[18, 19] It is also possible to classify with *rejection* or by dismissing the problematic (noisy) instances from the outset.[3, 18] However, from the point of view of concept learning, interfering with data is undesirable. This explains various efforts to improve on the random subsampling.[18, 19]

Existence of a class imbalance is usually ignored if the results are 'acceptable' or better.[15] In like situations the emphasis is on presentation metrics rather than addressing underlying causes of the underperformance. The balanced accuracy, or the arithmetic mean of sensitivity and specificity (half-sum of the two), gives an idea of the performance were the data class-balanced. As the imbalance increases, the sensitivity, as a rule, is lost to the specificity. In fact, the two are linearly related, parametric on the overall (irrespective of class affiliation) success rate.[20] The balanced accuracy is a single point estimate of area under the curve (AUC) on the *ROC* – "receiver operating characteristic" - plot (where 'y' is sensitivity and 'x' is one less specificity).[12] AUC is a measure of classifier responsiveness to a control, in continuous spaces usually quantifying position of the *decision surface*. Another measure of accuracy, invariant under the class imbalance change, is the geometric mean of sensitivity and specificity, their g-mean (square root of product of the two).

The ferocity of class imbalance largely depends on class data densities and how far the classes are removed from each other. That their shapes are also important is usually neglected in modelling.[20] In the meantime, any study is usually focused on a particular corner of data distribution. Without that, the notions of *data drift or concept shift* would probably be vacuous.[6] In diagnostic of chronic diseases there are parameters that stipulate the population to be screened, particularly the age.[7] Can this be that 'thoroughfare' for escaping from the class imbalance?

Feature selection is rarely discussed in the context of class imbalance.[10, 11, 23, 24] When a feature-set is selected by the backward elimination, in the current work mediated by the NB classifier, the effect of imbalance grows bolder as the feature-set becomes smaller. If the data was all-continuous and classification binary, this would be easy to interpret, since one should expect that the selected features to a greater extent than others are collinear with the axis connecting class centres, along which the class overlap is especially large. Narrowing the selection down deprives of a leeway of getting around the 'thicket'. Having more features thus helps to offset the class imbalance. Also, in this work the objective in feature subset selection was to elevate not the balanced accuracy but the smaller of sensitivity and specificity. There are features inherently in favour of the first or the second, such as for DM are the waist circumference to height ratio and the body mass index, respectively.[15]

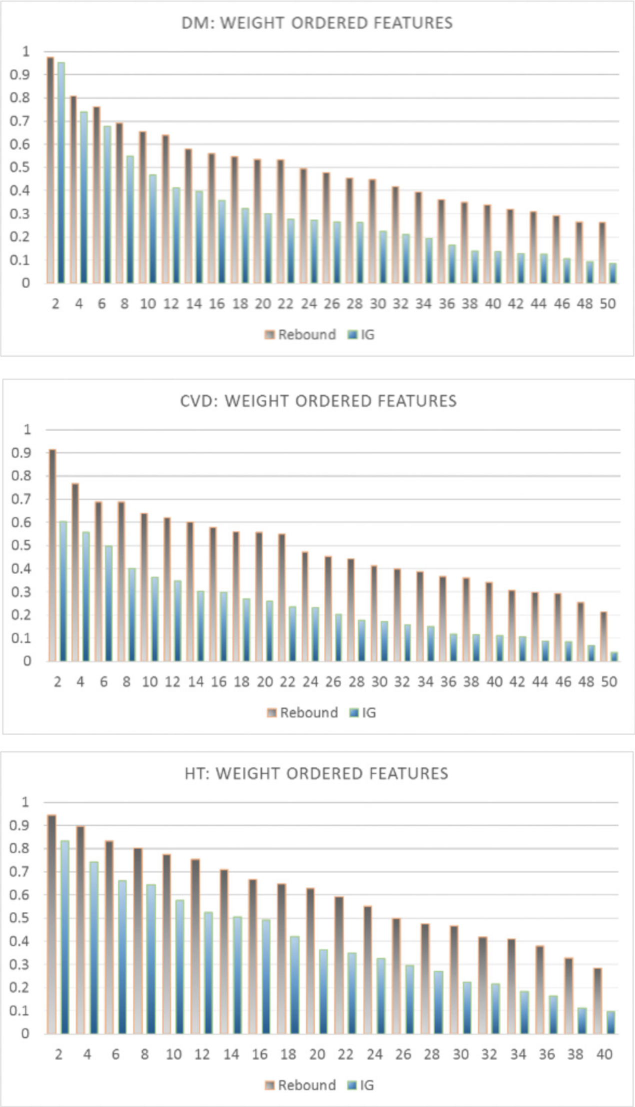

Figure 3. VDR (Rebound) and IG feature weights relative to the set-highest in diminishing order

5.4 IG vs VDR

IG-reliant k-NN clearly underperforms comparing with the weighting by VDR as seen from the results. In Figure 3 the weights on the basis of two measures are compared for the short feature-sets specific to DM, CVD and HT. The features are listed in diminishing order of weight, relative to the strongest feature on either list. Unlike their VDR counterparts, the IG weights decrease more rapidly and exhibit a concave pattern. This is attributable to the non-linear element in the design of IG (Eq.4). Evidently, by amplifying strong influences and playing down weak ones, IG antagonises the features to assist the agenda of their selection. Such a dissociation is not ideal at least for k-NN where the effect is achieved through a joint effort of the feature-set constituents.

However, the orderings from strongest to weakest are similar for IG and VDR. The sum of absolute differences in item position is a measure on a par to the Spearman's correlation coefficient for ranked lists.[25] Intuitively, the biggest divergence is attained if the ordering of the second list is opposite to the first. This is because repositioning from the fringe of a list is always able to gain more than from the middle thereof, and the opportunities wane with each move. Therefore, the characterisation of utter disarray is $r^2/2$ if r is even, $(r^2 - 1)/2$ if r is odd, where r is the lowest rank (length of the list). However, were it the maximum,[25] the solution is not unique. Compare, for example, the sequences: '7-6-5-4-3-2-1', '6-7-5-4-3-1-2' and '5-6-7-4-1-2-3' - they are all 24 positions off from the '1-2-3-4-5-6-7', not only the first. Having multiple solutions for the chaotic state does not preclude from using the disorder magnitude as a benchmark, though. Relative to it, the total shift in position for short-listed features is only 0.12 for DM, 0.09 for CVD and 0.08 for HT.

6. Conclusion

The motivation behind this research is that for patients on treatment it is difficult, if not impossible, to verify their status based on new data. Many medications specifically target the parameters involved in rules representing gold standards for diagnostic. On the face of it, this is done to avert complications, but deeply with a hope vested in the cyclic nature of metabolism. In diabetes the blood glucose is being watched, in hypertension - the blood pressure. For cardiovascular disease the diagnostic is circumstantial; nonetheless once a treatment is started the patient data becomes compromised. Normal levels of glucose, blood pressure or symptoms not presenting are quite possible, but this does not mean that disease causes had been addressed and patients are cured. However, the necessary information might be hidden in other parameters. The research conducted presently reaffirms the feasibility of identifying markers of disease other than those involved in metrics used for diagnostic of new patients. Also, it is inevitable that a large number of features is required. It is unlikely that any strong predictors could be missed out in a focused treatment. While it is possible to significantly reduce the number of variables, a variety of tests, measurements, examinations and history are available that can substitute for one another.

Patient data represents a high mix of different attribute types. Simultaneous handling of both continuous and nominal attributes is awkward. The conversion to discrete type is preferable as it offers a much more compact representation and consequently faster processing. However, not many methods are available for all-discrete domains. Any method can at least provide a basis for comparison. However, much more can be extracted from weak learners via *ensemble* techniques. In the current work quite a different angle was demonstrated - the limitation of the naive Bayesian turned into its advantage for feature-set selection. The approach universality was also evident. This is despite wrapping implies using the same method for both selection of features and classification.

The value distance metric coupled with the nearest neighbour approach to classification is a technique not being often quoted. In the meantime, VDM has some unique properties. Particularly, it holds the promise of assigning flexible weights to features that would be rigidly 'juxtaposing' otherwise. In the introductory discourse the conformity of VDM assigned weights and their range was given a due attention. During the evaluation VDM has demonstrated high-ranking results, although some aspects of them were less impressive. However, the overall results were not as good as by two other methods with rigid weights. While to a disappointment, this is in agreement with previous reports. Possible causes of the underperformance had been analysed. It was proposed that VDM may be vulnerable to the class imbalance. If so, this does not make VDM especially different from other methods and can be remedied.

The susceptibility of VDM to class imbalance is counterintuitive as it does not have corresponding terms in its formulation. For instance, NB has an explicit link to the majority rule. The information gain, used in this work for comparison as a weighting method for k-NN, also includes priors that would be affected by class imbalance. Although for VDM it did not work, a formulation not relying on prior probabilities is attractive with a view of class imbalance dependence reduction. This led to a VDM inspired weighting scheme for k-NN, albeit with rigid weights. The method, named the value displacement rebound, fared better than others on all counts among the four methods compared. So much so, its built-in ability to correct for class imbalance may have been helped by a tailored objective in the NB-based feature-set selection method.

References

[1] Stanfill C, Waltz D. Toward memory-based reasoning. Communications of the Association for Computing Machinery. 1986; 29: 1213-28. https://doi.org/10.1145/7902.7906

[2] Cost S, Salzberg S. A weighted nearest neighbor algorithm for learning with symbolic features. Machine Learning. 1993; 10: 57-78. https://doi.org/10.1007/BF00993481

[3] Stranieri A, Yatsko A, Golden I, et al. Capped k-NN editing in definition lacking environments. Journal of Pattern Recognition Research. 2013; 8(1): 39-58. https://doi.org/10.13176/11.465

[4] 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

[5] Domingos P. Context-sensitive feature selection for lazy learners. Artificial Intelligence Review. 1997; 11: 227-53. https://doi.org/10.1023/A:1006508722917

[6] Beringer J, Hullermeier E. An efficient algorithm for instance-based learning on data streams. In Perner P (editor): ICDM 2007; LNAI 4597:34-48. Springer-Verlag. https://doi.org/10.1007/978-3-540-73435-2_4

[7] American Diabetes Association. Standards of medical care in diabetes: Classification and diagnosis of diabetes. Diabetes Care. 2020; 43(S1): 14-31. https://doi.org/10.2337/dc20-S002

[8] Wilson DR, Martinez TR. Improved heterogeneous distance functions. Journal of Artificial Intelligence Research. 1997; 6: 1-34. https://doi.org/10.1613/jair.346

[9] Ting KM. Discretisation in lazy learning algorithms. Artificial Intelligence Review. 1997; 11: 157-74. https://doi.org/10.1023/A:1006504622008

[10] 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

[11] Aha DW. Feature weighting for lazy learning algorithms. In Liu H, Motoda H (editors) Feature Extraction, Construction and Selection: A Data Mining Perspective. 2001; 13-32. Kluwer.

[12] Kononenko I, Kukar M. Machine learning and data mining: introduction to principles and algorithms. 2007. Horwood. https://doi.org/10.1533/9780857099440

[13] Kononenko I. On bias in estimating multi-valued attributes. Proceedings of the 14-th International Joint Conference on Artificial Intelligence 1995; 1034-40. Morgan Kaufmann.

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

[15] 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

[16] Pazzani MJ. Searching for dependencies in Bayesian classifiers. In Fisher D, Lenz H-J (editors) Learning from Data: AI and Statistics V. 1996; 239-48. Springer-Verlag. https://doi.org/10.1007/978-1-4612-2404-4_23

[17] Pazzani MJ. Constructive induction of Cartesian product attributes. In Liu H, Motoda H (editors) Feature extraction, construction and selection: A data mining perspective; 2001. 341-54. Kluwer.

[18] Palacios AM, Sanchez L, Couso I. Equalizing imbalanced imprecise datasets for genetic fuzzy classifiers. International Journal of Computational Intelligence Systems. 2012; 5(2): 276-96. https://doi.org/10.1080/18756891.2012.685292

[19] Gui C. Analysis of imbalanced data set problem: The case of churn prediction for telecommunication. Artificial Intelligence Research. 2017; 6(2): 93-9. https://doi.org/10.5430/air.v6n2p93

[20] Tasinaffo PM, Gonsalves GS, da-Cunha AM, et al. Using Monte Carlo method to estimate the behavior of neural training between balanced and unbalanced data in classification of patterns. Artificial Intelligence Research. 2018; 7(2): 1-25. https://doi.org/10.5430/air.v7n2p1

[21] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 1997; 97(1/2): 273-324. https://doi.org/10.1016/S0004-3702(97)00043-X

[22] Kohavi R, John GH. The wrapper approach. In Liu H, Motoda H (editors) Feature extraction, construction and selection: A data mining perspective 2001, 32-50. Kluwer.

[23] Liu H, Motoda H (editors). Computational methods of feature selection 2008; Chapman & Hall / CRC.

[24] Huang SH. Supervised feature selection: a tutorial. Artificial Intelligence Research. 2015; 4(2): 22-37. https://doi.org/10.5430/air.v4n2p22

[25] Bhamidipati NL, Pal SK. Comparing rank-inducing scoring systems. In Proceedings of the 18-th International Conference on Pattern Recognition (ICPR). 2006; 1-4. IEEE. https://doi.org/10.1109/ICPR.2006.390