Ultimate Order Statistics-Based Prototype Reduction Schemes

A. Thomas and B. John Oommen*

School of Computer Science, Carleton University, Ottawa, Canada, K1S 5B6
smithasam2007@yahoo.com, oommen@scs.carleton.ca

Abstract. The objective of Prototype Reduction Schemes (PRSs) and Border Identification (BI) algorithms is to reduce the number of training vectors, while simultaneously attempting to guarantee that the classifier built on the reduced design set performs as well, or nearly as well, as the classifier built on the original design set. In this paper, we shall push the limit on the field of PRSs to see if we can obtain a classification accuracy comparable to the optimal, by condensing the information in the data set into a *single training* point. We, indeed, demonstrate that such PRSs exist and are attainable, and show that the design and implementation of such schemes work with the recently-introduced paradigm of Order Statistics (OS)-based classifiers. These classifiers, referred to as Classification by Moments of Order Statistics (CMOS) is essentially anti-Bayesian in its *modus operandus*. In this paper, we demonstrate the power and potential of CMOS to yield single-element PRSs which are either "selective" or "creative", where in each case we resort to a non-parametric or a parametric paradigm respectively. We also report a single-feature single-element creative PRS. All of these solutions have been used to achieve classification for real-life data sets from the UCI Machine Learning Repository, where we have followed an approach that is similar to the Naïve-Bayes' (NB) strategy although it is essentially of an anti-Naïve-Bayes' paradigm. The amazing facet of this approach is that the training set can be reduced to *a single* pattern from each of the classes which is, in turn, determined by the CMOS features. It is even more fascinating to see that the scheme can be rendered operational by using the information in a *single feature* of such a *single data point*. In each of these cases, the accuracy of the proposed PRS-based approach is very close to the optimal Bayes' bound and is almost comparable to that of the SVM.

Keywords: Prototype Reduction Schemes, Classification using Order Statistics (OS), Moments of OS.

1 Introduction

In traditional non-parametric classification, the training patterns play a significant role in the classification process. This is because a decision boundary is obtained by considering *all* the samples in the training set. However, modern

* *Chancellor's Professor*; *Fellow: IEEE* and *Fellow: IAPR*. This author is also an *Adjunct Professor* with the University of Agder in Grimstad, Norway.

S. Cranefield and A. Nayak (Eds.): AI 2013, LNAI 8272, pp. 421–433, 2013.

rapid advancements in this field have led to the development of efficient classification methods in which the schemes achieve the classification based on a *subset* of the training patterns. A Prototype Reduction Schemes (PRS) is a generic method for reducing the number of training vectors, without affecting the performance of the classifier built on the reduced design set [1–4]. Instead of considering all the training patterns for the classification, a subset of the whole set is selected based on certain criteria. The training is then performed on this reduced set, which is also called the "Reference" set. More recent advances have involved the use of Border Identification (BI) algorithms [5–8] to choose these prototypes from the so-called "border" points of the various classes.

Traditionally, a good PRS can reduce the size of the training set to a small percentage (for example, 10%) of the original set. But how small can one make this reduced set? Is it possible to, at least conceptually, reduce the set of prototypes to contain *only a single element* from each class. The aim of this paper is to investigate this issue both conceptually and from a practical perspective. Indeed, we shall demonstrate that we can push and attain the limit on the field of PRSs to obtain a classification accuracy comparable to the optimal, by condensing the information in the data set into a *single training* point. Apart from showing that such a PRS exists and is attainable, we shall also show that the design and implementation of such a mechanism relies on the recently-introduced paradigm of Order Statistics (OS)-based classifiers.

One should, of course, mention that the new point obtained by invoking the PRS is not necessarily a member of the original data set. Rather, it can be an artificially created point, representative of the training set, as perceived from the perspective of the data sets OSs.

We now consider another facet of a typical PRS-based PR solution. Whenever a practitioner designs a PRS, he works with the premise that *all* features are crucial for the classification. The problem that is "dual" to the PRS problem is the following: Apart from reducing the size of the "Reference" set, is it possible to also reduce the number of features utilized within the latter. This paper addresses both of these issues simultaneously. To be specific, we state that the OS-based PRS scheme that we propose has the fascinating property that it can be rendered operational by using the information in a *single feature* of the *single data point* obtained using an OS-based computation. Indeed, in each of these cases, the accuracy of this approach is very close to the optimal Bayes' bound and is almost comparable to that of the SVM. In a nutshell, this is the fundamental contribution of this paper, and we are not aware of any reported comparable results.

To put this paper in the right context, a word about these OS-based classifiers is not out of place [9–11]. Almost all the well-known classifiers involved in pattern classification are based on a Bayesian principle which aims to maximize the *a posteriori* probability, where they have been characterized by their respective indicators such as their means, variances etc.. In the field of PR, however, there are some families of indicators that have noticeably been uninvestigated, specifically those related to its Order Statistics (OS). The interesting point about these indicators is that some of them are quite unrelated to the traditional

moments themselves, and in spite of this, have not been used in achieving PR. The main question that has earlier excited our interest is whether these indicators/indices possess any potential in PR.

The salient differences between the traditional Bayesian paradigm and the newly-proposed OS-based anti-Bayesian paradigm can be highlighted as below. Consider Figure 1, where for simplicity, we have used unit-lengthed intervals to display the span of the two class-conditional distributions. Whenever a testing sample comes from these distributions, the CMOS will compare the testing sample with the *higher*-order 2-OS, $E[\mathbf{x}_{2,2}]$ of the first distribution, i.e., $\frac{2}{3}$, and with with the *lower*-order 2-OS $E[\mathbf{x}_{1,2}]$ of the second distribution, i.e., $h + \frac{1}{3}$, and the sample will be labeled with respect to the class which minimizes the corresponding quantity, as shown in Figure 1. We emphasize that the comparison is not made with the *means* of the two distributions, but with certain non-central outlier-like points, rendering it "Anti"-Bayesian. Observe that for the above rule to work, we must enforce the ordering of the OS of the two distributions, and this requires that $\frac{2}{3} < h + \frac{1}{3} \implies h > \frac{1}{3}$. The case when this condition is not satisfied, and the details of CMOS have been explained in [9–11].

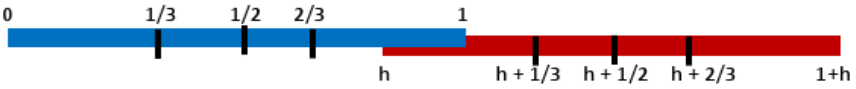

Fig. 1. A schematic of OS-based Anti-Bayesian Classification

This paper takes this concept to the next level, i.e., to that concerning PRSs.

From an overall perspective, we now discuss how we are to achieve our goal to reduce the cardinality of the OS-based PRS to be unity for each class. First of all, we know that PRSs can be broadly classified as being "selective" or "creative" [12]. A "selective" PRS yields as its output a set of prototypes which are *chosen* from the original training points. As opposed to this, a "creative" PRS *creates* a set of artificial points which may not be found in the original training set, and these points are thereafter used in the classification.

We first study the task of designing "selective" OS-based PRSs in Section 4. Since, at this juncture, we are not willing to assume a distributional form for the corresponding features, we are forced to work with the non-parametric representation that the training data captures. By working with the multi-dimensional non-parametric form of the data, and by thereafter invoking an OS-based paradigm, we are able to obtain a *single* prototype with which we can accomplish efficient classification. This *single* prototype is, as a vector, a "created" point, although, in every single dimension, the value is "selected" from the actual training sample that is closest to the value specified by the OS value.

Two versions of this strategy have been proposed, namely, the first which considers the entire vectorial form of the resultant prototype (in Section 4.1), and the second which invokes a majority vote by considering the OS-based classification of the individual features. The latter, which is a *Scalar*-Based Selective PRS, has been described in Section 4.2. It is worth mentioning that

the classification results obtained by both these methods – both of which involve only a *single* prototype – are quite satisfactory, and are comparable, though understandably, marginally inferior, to those obtained from a NB or SVM strategy.

After investigating selective PRSs, we subsequently consider the task of designing "creative" OS-based PRSs in Section 5. In this case, we assume a distributional form for the corresponding features, and so we proceed to work with the parametric representation that the training data captures. By working with a multi-dimensional parametric form of the data, and by thereafter invoking an OS-based paradigm, we succeed in obtaining a *single* prototype in the "Reference" set, which can be used for classification. This process has been explained in Section 5.1. As in the non-parametric case, we have also developed a *Scalar*-Based Creative PRS in Section 5.2. Again, it is worth mentioning that the classification results obtained from both these parametric strategies (i.e., the vector, and the majority-voted individual-feature based) are quite satisfactory, and comparable, though marginally inferior, to those obtained from a NB or SVM strategy.

The final concluding contribution is actually far more ambitious. It consists of using only a *single* feature of a *single* prototype. In this case, in Section 6, we have designed a "creative" PRS scheme which merely includes the OS-based points of a single feature, where the $\frac{n-k+1}{n+1}^{th}$ percentile of *this* feature of the first class, and the $\frac{k}{n+1}^{th}$ percentile of *this* feature of the second class, are the corresponding "prototypes". It is clear that the accuracy of this *scalar*-based OS will be inferior to that of the corresponding *vector*-based OS. However, astonishingly enough, the accuracy does not degrade significantly – the resultant classifier still yields an accuracy that is acceptable considering the fact that one requires only a single *scalar* comparison to achieve the classification.

The reader must observe that the intent of this paper is not to compare the resultant classification accuracies with those obtained from an entire ensemble of classification methodologies. Rather, our aim is to show that we can obtain very efficient classification by merely using a single (vector or scalar) prototype which is either selected or created. Thus, we have compared our proposed scheme with only *three* standard algorithms which have been universally considered as benchmarks. We believe that the results presented here conclusively demonstrate the power of our contribution.

1.1 Contributions of This Paper

The novel contributions of this paper are:

- We propose a "selective" PRS which can be metaphorically perceived to be the "Ultimate" selective PRS because, by using a non-parametric paradigm, it reduces the size of the "Reference" set to be a *single* pattern from each class, which is thereafter utilized in the classification;
- We also propose a "creative" PRS which can be considered to be the "Ultimate" creative PRS because, by invoking a parametric paradigm, it also reduces the size of the "Reference" set to be a *single* pattern from each class;
- In both of the above cases, we have also shown that it is possible to derive a majority-based PRS which fuses the classification results of the various

features of the *single* d-dimensional prototype. The classification accuracies of these fused scalar schemes are marginally worse than those of the corresponding vector-based algorithms;
- We have also shown that it is possible to derive a single scalar prototype, i.e., one which involves only a a *single* feature of a *single* d-dimensional vector. The classification accuracy of this single-scalar PRS is marginally worse than that of the vector-based methods;
- In every case, we demonstrate, by testing the algorithms on real-life data sets from the UCI repository, that the new PRS-based classification schemes yield accuracies comparable to the traditional NB classifiers, and even the SVM, even though the computations needed are, really, of an atomic magnitude.

In the interest of space, the formal algorithms for all these strategies cannot be included here. But in the interest of completeness, as a representative example, we have included the formal algorithm for for one of these strategies, namely for the Ultimate *Vector*-based Creative PRS in Section 5.1.

We conclude this section by remarking that, to the best of our knowledge, analogous results have been unreported in the literature.

2 CMOS-Based Classification: The Generic Classifier

The multi-dimensional OS-based classifier is based on its uni-dimensional counterpart developed earlier. Since its understanding is crucial to this paper, it is briefly explained here.

Consider a 2-class problem with classes ω_1 and ω_2, where their class-conditional densities are $f_1(x)$ and $f_2(x)$ respectively (i.e, their corresponding distributions are $F_1(x)$ and $F_2(x)$ respectively). If we perform a classification based on ν_1 and ν_2, the *medians* of the distributions, this is equivalent to the strategy in which the task is performed based on a *single* OS. For all symmetric distributions, this classification accuracy attains the Bayes' accuracy – which is not too astonishing because the median is identical to the mean. But the intriguing aspect emerges when we use higher order OS that are not located centrally (close to the means), but rather *distant* from the means. Indeed, for uni-dimensional OS-based PR, our methodology is based on considering the n-order OSs, and comparing the testing sample with the $n-k$ OS of the first distribution and the k^{th} OS of the second. By considering the entire spectrum of the possible values of k, the results in and showed that the specific value of k is usually not so crucial. Further, if these symmetric pairs of the OS are used in PR, the classification based on *these* attains the optimal Bayes' bound for a large number of symmetric distributions of the exponential family. The PR is near-optimal when the distributions are asymmetric.

Theses results were generalized for multi-dimensional distributions by invoking a Naïve-Bayes' approach, which essentially implies that that the first moments of the OS in each of the dimensions are uncorrelated.

With this as the background, we shall now consider how we can derive single-element OS-based PRSs which can be used to design classifiers for real-life data. Since our solutions have been tested on both artificial and real-life data-sets, we shall, in the interest of continuity, briefly describe the sets that we have used.

3 Experimental Data Sets

3.1 Artificial Data Sets

For a *prima facie* testing of artificial data, we generated two classes that obeyed Gaussian distributions. To do this, we made use of a Uniform $(0, 1)$ random variable generator to generate data values that follow a Gaussian distribution. The expression $\mathbf{z} = \sqrt{-2ln(u_1)}\ cos(2\pi u_2)$ is known to yield data values that follow $N(0, 1)$ [13]. Thereafter, by using the technique described in [14], one can generate Gaussian random vectors which possess any arbitrary mean and covariance matrix. The means of the classes were $[2\ \ 2\ \ 2\ \ 2\ \ 2]^T$ and $[-2\ \ -2\ \ -2\ \ -2\ \ -2]^T$ respectively, and the covariances of the two classes were identical and had the form[1]:

$$\Sigma = \begin{bmatrix} a^2 & b & 0 & a & \alpha ab \\ b & 2a+3b & 0 & b & a \\ 0 & 0 & 1 & 0 & 0 \\ a & b & 0 & 2a+3b & b \\ \alpha ab & a & 0 & b & b^2 \end{bmatrix}$$

This rendered the classes to have an optimal linear classifier. With regard to the cardinality of the data set, each of the classes had 200 instances in the corresponding 5-dimensional space.

3.2 Real-Life Setup

The data sets [15] used in this study have two classes, and the number of attributes varies from 4 up to 32. The data sets are given in Table 1.

4 OS-Based "Selective" PRSs Using a Non-parametric Perspective

In this section, we discuss the problem of designing a "Selective" OS-based PRS. Since we are ultimately going to select a training sample, at this juncture, we take the position that we are not willing to assume a *distributional form* for the corresponding features. Consequently, we are forced to work with the non-parametric representation that the training data captures. This implies that one has to resort to a non-parametric avenue in which we are able to compute the corresponding prototypes by approximating the distribution using a multi-dimensional kernel. Although a generalized kernel could be used for this phase, in the interest of simplicity, for a *prima facie* case, we have opted to use a simplistic bin-based approach. Once the histogram of the features has been obtained in each dimension, the training sample that lies closest to the point representing the $\frac{n-k+1}{n+1}^{th}$ percentile of the first distribution and the $\frac{k}{n+1}^{th}$ percentile of the

[1] In our experiments, we set $a = 5$, $b = 4$, and $\alpha = 0.4$.

Table 1. The Real-life data sets used in our experiments, where C, I and R represent Categorical, Integer and Real Respectively

Data set	No. Instances	No. Attributes	No. Classes	Attribute Type
WOBC	699	9	2	I
WDBC	569	32	2	Real
WDBC	569	32	2	R
Diabetes	768	8	2	I, R
Hepatitis	155	19	2	C, I, R
Iris	150	4	3	Real
Mushroom	8124	22	2	C
Statlog (Heart)	270	13	2	C, R
Statlog (Australian Credit)	690	14	2	C, I, R
Vote	435	16	2	C, I

second distribution of the given data sets is *selected* to be the prototype of interest. Indeed, by using these *selected* patterns as vector prototypes – *a single one from each class* – one can now achieve classification. One should observe that this *single* prototype is, as a vector, a "created" point, although, in every single dimension, the value is "selected" from the actual training sample that is closest to the value specified by the OS value.

Although the specific value of k is not so crucial [9–11], in this paper, as mentioned earlier, we have set $k = 1$, implying that we have, in each dimension, worked with the pattern that falls at the $\frac{2}{3}$ percentile of the first distribution and the pattern that falls at the $\frac{1}{3}$ percentile of the second.

To obtain the final PRS, we can envision two methodologies, namely where the computations are vector-based or scalar-based, which are described below.

4.1 The *Vector*-Based Selective OS-Based PRS

The *Vector*-based selective OS-based PRS is obtained by comparing the testing sample with the prototype procured by the above process. Such a comparison can be achieved using any metric, but for the sake of simplicity, we have utilized the well-known Euclidean norm.

The proposed method has been rigorously tested on the various artificial and real-life data sets obtained from the UCI repository [15] described above. It has also been compared with other well-known schemes including the NB, SVM, and the kNN. In order to obtain the results, the algorithms were executed 50 times with the 10-fold cross validation scheme. The results are tabulated in Table 2. To ensure standardization , the performance of the benchmark classifiers are taken from [16–18]. By examining the table of results (see Column 6), we can see that the proposed algorithm can achieve a comparable classification when compared to the other traditional classifiers, which is particularly impressive because once the *single* prototype has been computed after the training phase, the testing is done by exactly two vector-based computations (one for each class), comparing the testing sample with the resultant prototypes. For example, for

the Breast Cancer data set, we can see that the new approach yielded a accuracy of 95.06% which should be compared to the accuracies of the SVM (96.99%), NB (96.40%) and the kNN (96.60%). The reader will observe that the classification accuracies for all the data sets is commendable except for the "Diabetes" set. This is because, for this data set, the approximation of the distributions using simplistic histograms in the d-dimensional space is rather crude. Superior results are obtained in this case when we resort to obtaining the OS-based points using the criteria explained in Section 5.1.

4.2 The *Scalar*-Based Selective OS-Based PRS

In the *Scalar*-based selective OS-based PRS, the patterns are treated as a group of scalars and a classification is performed for each dimension. Thereafter, the final determination of the identity of the testing sample is achieved based on a majority vote. The scalar-based selective CMOS has been tested on the various artificial and real-life data sets and the results are tabulated in Table 2. If we examine the table (see Column 8), one can see that the approach yields a near optimal accuracy for the all the data sets except the Diabetes data set, which, as before has a poor accuracy for all the classifiers, and for which the histogram leads to a very crude approximation. For example, if we consider the Hepatitis data set, the proposed approach yields an accuracy of 81% while the traditional classifiers yields 84.54% (SVM), 82.58% (NN) and 83.19% (NB), which is still quite astonishing considering that all the information in the entire training set has been crystallized into a single prototype *distant from the mean*.

We now move on to present the vector and scalar-based "Creative" PRSs in which the Reference set has only a single element.

5 A CMOS-Based "Creative" PRS Using a Parametric Perspective

We now consider the task of designing a "creative" OS-based PRS, where we again aim to attain the goal that the cardinality of the Reference set is unity. Since we are now willing to permit the option of assuming a distributional form for the corresponding features, we have chosen to resolve this fundamental issue by invoking a strategy analogous to a Naïve-Bayes' approach, although it, really, is of an *anti*-Naïve-Bayes' paradigm. As a Naïve-Bayes' strategy requires the uncorrelation of the features, if we consider a k-OS CMOS, we need to determine, for every feature, the $\frac{n-k+1}{n+1}^{th}$ percentile of the first distribution and the $\frac{k}{n+1}^{th}$ percentile of the second distribution. From an anti-Naïve-Bayes'perspective, we can obtain the corresponding values of all of the features by assuming a Gaussian[2] distribution for all the features. The OS-based PRS that we thus propose

[2] Any other member of the exponential family described in [9] could have just as well been used. We have chosen to use the Gaussian distribution because it is more general than the others, and involves the means and the variances of the features.

here again consists of the *single created* point in the d-dimensional space characterized by the location of the $\frac{n-k+1}{n+1}^{th}$ percentile of the first distribution and the $\frac{k}{n+1}^{th}$ percentile of the second distribution. As shown in [9], for the value of $k = 1$, the 2-OS CMOS positions for the classes that follow a Gaussian distributions can be expressed as $u_1 = \mu_1 - \frac{\sigma}{\sqrt{2\pi}}$ and $u_2 = \mu_2 + \frac{\sigma}{\sqrt{2\pi}}$. We thus opt to use these expressions to obtain the corresponding CMOS positions, whence the vector and scalar-based PRS schemes are derived.

5.1 The *Vector*-Based "Creative" OS-Based PRS

For this approach also, we consider the possibility of perceiving the training set as vectors or as scalars. The *Vector*-based "Creative" OS-based PRS considers the final prototype as a vector, which has been artificially created as a new pattern by resorting to the expressions for u_1 and u_2. The testing sample is then compared with the *single* OS-based prototype, and the identity is determined with regard to how distant it is from the latter. Since the individual variances are known, this distance is computed using the Mahalanobis distance.

The formal algorithm for this approach is given in Algorithm 1.

Algorithm 1. `Vector_based_Creative_PRS`(T, TP)

Input:

T: The training set, comprising of elements T_1 and T_2 from classes ω_1 and ω_2 respectively.
TP: the testing set

Output:

Classification for TP

Method:
Training
for i = 1 to d **do**
  Estimate mean of T_1 as μ_{1i} and mean of T_2 as μ_{2i}
  Estimate the standard deviations of T_1 and T_2 as σ_{1i} and σ_{2i}
end for
for i = 1 to d **do**
  Determine the i^{th} component of $\mathbf{u_1}$, $u_{1i} = \mu_{1i} - \frac{\sigma_{1i}}{\sqrt{2\pi}}$
  Determine the i^{th} component of $\mathbf{u_2}$, $u_{2i} = \mu_{2i} + \frac{\sigma_{2i}}{\sqrt{2\pi}}$
end for
End_Training

Testing
for all $\mathbf{x} \in TP$ **do**
  if $M\_Dist(\mathbf{u_1}, \mathbf{x}) < M\_Dist(\mathbf{u_2}, \mathbf{x})$ **then**
    Assign $\mathbf{x}$ to class ω_1
  else
    Assign $\mathbf{x}$ to class ω_2
  end if
end for

End_Testing

End Algorithm

Table 2. Classification of Real-life data sets by CMOS

Data set	Traditional Classifiers			CMOS Classifier			
	NB	NN	SVM	Vector		Scalar	
				Creative	Selective	Creative	Selective
WOBC	96.40	96.60	96.99	96.94	95.06	94.35	92.06
WDBC	92.97	96.66	97.71	93.43	90.07	89.25	86.82
Diabetes	73.11	71.90	73.84	73.76	65.74	76.74	43.41
Hepatitis	83.19	82.58	84.54	76.67	75.13	81.87	81.00
Iris	95.13	96.00	96.67	94.4	92.50	93.80	77.80
AU Credit	87.40	85.90	85.51	94.76	84.21	83.03	48.19
Heart	83.00	84.40	85.60	84.59	83.93	77.11	60.67
Vote	94.29	90.23	94.33	93.43	91.0	89.10	85.36

The vector-based *Creative* CMOS has been tested for the same data sets as before, and the results are tabulated in Table 2. From the table (see Column 5), we can conclude that the new approach is comparable with the other well-used and well-established classifiers. This approach achieves "almost" optimal classification when compared to the traditional classifiers. For example, if we consider the classification of the Breast Cancer data set, we see that Algorithm achieves 96.94% accuracy as opposed to the 96.99% of SVM, 96.40% of NB and 96.6% of NN. One can see that the difference in the accuracies is almost negligible. For the other data sets too, this approach attains a near-optimal classification when compared to the traditional classifiers, even though there is only a single element in the Reference set, and the testing involves only two vector comparisons.

5.2 The *Scalar*-Based "Creative" OS-Based PRS

In this approach, each pattern was considered as a vector, and the distance calculations were based on the Mahalanobis metric. As in the case of the selective scheme described in Section 4.2, a similar classification can be achieved by considering the various feature values as scalars and by accomplishing the task by computing the majority vote.

The scalar-based creative CMOS has also been tested on the various artificial and real-life data sets and the results are tabulated in Table 2 (see Column 7). Again, an examination of the table shows that the classification results are near-optimal. For example, if we consider the Vote data set, the proposed approach yields an accuracy of 93.43% while the traditional classifiers yields 94.33% (SVM), 90.24% (NN) and 94.29% (NB). Observe that the prototype-based NN performs even better than the traditional NN which involves the entire training set, which is quite astonishing considering that all the information in the entire training set has been crystallized into a single newly-created prototype.

6 Classification Based on One Selected Feature

In this section we have embarked on an even far more ambitious goal which consists of seeing if we could do the classification by using only a *single* feature

of a *single* prototype. To achieve this goal, we have operated with the philosophy proposed in Section 5 and designed a "creative" vector PRS. But rather than use all the components of the vector in the classification, we have merely chosen the OS-based points of a *single feature*, where the $\frac{n-k+1}{n+1}^{th}$ percentile of *this* feature of the first class, and the $\frac{k}{n+1}^{th}$ percentile of *this* feature of the second class, are the corresponding "prototypes" (where we have, as usual, used $k = 1$).

The proposed approach of has been tested on the artificial and real-life data sets described earlier, and the results are tabulated in Table 3. If we closely investigate the table, one can see that the method attains a comparable classification when compared to the traditional classifiers. Specifically, for the Diabetes data set, if the classification is performed based on the OS positions of the feature *Plasma Glucose Concentration*, an accuracy of 73.63% is attained as opposed to the accuracy of 73.84% attained by SVM . The reader should not be surprised that the accuracies are not always so outstanding. However, astonishingly enough, the accuracy does not degrade significantly – the resultant classifier still yields an accuracy that is acceptable considering the fact that one requires only two *scalar* comparisons to achieve the classification.

Table 3. Classification of Artificial and Real-life data sets using the Scalar-based *Creative* CMOS involving only a single dimension

Data set	**SVM**	**Dimension**	**Feature**	**CMOS**
Artificial Set	98.75	3	A3	98.475
WOBC	96.99	2	Uniformity of Cell Size	93.04
WDBC	97.71	27	Worst Compactness	91.29
Diabetes	73.84	2	Plasma Glucose Concentration	73.63
Hepatitis	84.54	12	Ascites	83.93
Iris	96.67	4	Petal Width	95.5
AU Credit (Statlog)	92.1	7	A9	84.84
Heart (Statlog)	85.60	2	Chest Pain Type	78.52
Vote	94.33	4	Physician-fee-freeze	95.40

7 Conclusions

Almost all the well-known classifiers involved in pattern classification are based on a Bayesian principle which aims to maximize the *a posteriori* probability. Quite recently, a new paradigm, known as CMOS, the classification by moments of Order Statistics, has been introduced to attain the same task, but with a counter-intuitive philosophy as compared to the Bayesian principle. In [10], the foundational theory of the CMOS was introduced, and a generic classifier that can be used for any distribution was provided. The applications of CMOS on various uni-dimensional distributions of the exponential family were included in [9]. The results of [9] were extended for multi-dimensional distributions in [11].

In this paper, we have demonstrated the power and potential of CMOS to yield single-element PRSs which are either "selective" or "creative", where in each case we resort to a non-parametric or a parametric paradigm respectively. We have derived a single-feature single-element creative PRS. All of these solutions have been used to achieve classification for artificial and real-life data sets from the UCI Machine Learning Repository. All of the reported algorithms yield an acceptable accuracy when compared to many of the established benchmark methods. It is even more fascinating to see that our paradigm performs favorably by using the information in a *single feature* of such a *single data point*.

References

1. Garcia, S., Derrac, J., Cano, J.R., Herrera, F.: Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(3), 417–435 (2012)
2. `http://sci2s.ugr.es/pr/` (April 18, 2013)
3. Kim, S., Oommen, B.J.: On Using Prototype Reduction Schemes and Classifier Fusion Strategies to Optimize Kernel-Based Nonlinear Subspace Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 455–460 (2005)
4. Triguero, I., Derrac, J., Garcia, S., Herrera, F.: A Taxonomy and Experimental Study on Prototype Generation for Nearest Neighbor Classification. IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews 42, 86–100 (2012)
5. Duch, W.: Similarity Based Methods: A General Framework for Classification, Approximation and Association. Control and Cybernetics 29(4), 937–968 (2000)
6. Foody, G.M.: Issues in Training Set Selection and Refinement for Classification by a Feedforward Neural Network. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, pp. 409–411 (1998)
7. Foody, G.M.: The Significance of Border Training Patterns in Classification by a Feedforward Neural Network using Back Propogation Learning. International Journal of Remote Sensing 20(18), 3549–3562 (1999)
8. Li, G., Japkowicz, N., Stocki, T.J., Ungar, R.K.: Full Border Identification for Reduction of Training Sets. In: Bergler, S. (ed.) Canadian AI. LNCS (LNAI), vol. 5032, pp. 203–215. Springer, Heidelberg (2008)
9. Oommen, B.J., Thomas, A.: Optimal Order Statistics-based "Anti-Bayesian" Parametric Pattern Classification for the Exponential Family. Pattern Recognition (2013) (accepted for Publication)
10. Thomas, A., Oommen, B.J.: The Fundamental Theory of Optimal "Anti-Bayesian" Parametric Pattern Classification Using Order Statistics Criteria. Pattern Recognition 46, 376–388 (2013)
11. Thomas, A., Oommen, B.J.: Order Statistics-based Parametric Classification for Multi-dimensional Distributions (submitted for publication 2013)
12. Kim, S., Oommen, B.J.: A brief Taxonomy and Ranking of Creative Prototype Reduction Schemes. Pattern Analysis and Applications 6, 232–244 (2003)
13. Devroye, L.: Non-Uniform Random Variate Generation. Springer, New York (1986)
14. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, San Diego (1990)
15. Frank, A., Asuncion, A.: UCI Machine Learning Repository (2010), `http://archive.ics.uci.edu/ml` (April 18, 2013)

16. http://www.is.umk.pl/projects/datasets.html (April 18, 2013)
17. Karegowda, A.G., Jayaram, M.A., Manjunath, A.S.: Cascading K-means Clustering and k-Nearest Neighbor Classifier for Categorization of Diabetic Patients. International Journal of Engineering and Advanced Technonlogy 01, 147–151 (2012)
18. Salama, G.I., Abdelhalim, M.B., Elghany Zeid, M.A.: Breast Cancer Diagnosis on Three Different Datasets using Multi-classifiers. International Journal of Computer and Information Technology 01, 36–43 (2012)