THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

A Novel Feature Selection Algorithm for Strongly Correlated Attributes Using Two-Dimensional Discriminant Rules

Taufik DJATNA[†] Yasuhiko MORIMOTO[‡]

[†, ‡] Dept. Of Information Engineering, Graduate School of Engineering, Hiroshima University
1-7-1 Kagamiyama Higashi Hiroshima

739-8521, Japan

E-mail: [†]taufikdjatna@hiroshima-u.ac.jp, [‡]morimoto@mis.hiroshima-u.ac.jp

Abstract. Considerable attention has been devoted to the development of feature selection algorithms for various applications in the last decade. Most of them concentrate to the single attributes. In contrast, limited research work has been devoted to determine correlated and pairwise attributes or features due to the difficulty of the problem. We present a novel feature selection algorithm for strongly correlated attributes using the two-dimensional discriminant rules. We discover a subset of pairwise attributes whose target class is influenced not only by a single cause in numeric-based datasets, both categorical and continuous target classes. Our algorithm uses x-monotone optimization for determination of optimal region within datasets. For selection strategy, the region evaluation has been applied using skewness and kurtosis metric. The results are then compared to other numeric based attribute selection algorithms. The result shows a unique capability to reveal the importance of pairwise strongly correlated attributes that conventional methods missed to explore.

Keyword strongly correlated attributes, two-dimensional discriminant rules approach

1. Introduction

There are pervasive data with strongly correlated attributes in real world that related to classification and regression operation. For instance, it is natural to check all the features relate to the disorder occurrence of patient's weight and his blood pressure while we consider a diabetes status. Thus, in order to decide the most relevant pairwise attributes in one such cases, we will take into account a considerable amount of pair of attributes comparison.

Attributes selection is an important technique usually deployed for warranting the estimated accurately classification which can be obtained from any size of training data examples [1, 2]. Existing conventional attribute selection algorithms assume that there is no interdependency between attributes that warrant the process to decide a target class in a dataset [1-3]. Conventional approaches remove the correlated attributes as they will ruin the classification result. Contrary to this assumption, interdependency between features or attributes within any real world datasets are naturally occurred and applied in many fields of science. In this paper we propose a solution for those strongly correlated attributes selection by constructing an algorithm based on two dimensional discriminant rules approach [4-7] using the x-monotone optimization. Our main contributions to these problems as follows:

- We solve the pairwise attributes selection by investigating the strongly correlated features with nonlinearity into account.
- We construct a mechanism for the pairwise attributes selection strategy of the optimized x-monotone region on both categorical and continuous target class determination.

The novelty of this work is on the filtering the pairwise attributes using x-monotone region optimization both in classification and regression cases. For the separability criterion, we utilize the information gain metric within classification case and the interclass variance metric in regression. Both cases are included with selection strategy to determine the pair of correlated attributes

which are informative features characterizing the target class. These results are absent in the most conventional (single) feature selection algorithms. The results are significantly helpful to answer the interdependency between features which may occur in any datasets.

This paper is organized as follows: in Section 2 we discuss briefly previous work on related techniques. Section 3 briefly give the definitions and the framework of our attribute selection algorithm using 2-dimensional discriminant rule approach, followed by Section 4, where we discuss our pairwise attributes selection algorithm in detail. Section 5 presents and describes the implementation and experiments using real world data. Section 6 discusses a comparison to the single numeric attribute selection algorithms including their possible application in related field of real domain cases. Section 7 concludes the paper by summarizing our main contribution.

2. Related Work

Several feature selection techniques have been proposed in the literature, including some important survey on feature selection algorithms such Molina et al. [8] then Guyon and Elisseeff [9]. Many researchers involved in studying various important point of feature selection, such as the goodness of a feature subset in determining an optimal one [10]. Different feature selection can be categorized into wrapper model [11], filter [12-16] and embedded [1, 12]. The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subset. The most important point about this method is the higher computational cost [11]. The filter model separates feature selection from classifier learning and selects feature subsets that are independent of any learning algorithm. It relies on various measures of the general characteristics of training data such as distance, information dependency and consistency [10]. According to the availability of class labels, there are feature selection methods for supervised learning [16, 17] as well as for unsupervised learning [18, 19]. Existing feature selection methods mainly exploit two approaches: individual feature evaluation and subset evaluation [9]. Individual evaluation methods rank features according to their importance in differentiating instances of different classes and can only remove irrelevant features as redundant features likely similar rankings. Methods of subset evaluation search for a minimum subset of features that satisfies some goodness measure and can remove irrelevant features as well as redundant ones.

Embedded feature selections are associated with the induction algorithm. For example, some feature selection algorithms are embedded with other strong induction learning methods such as genetic algorithms and neural network using method such as Markov Blanket [20].

There are some proposals on how to find important optimal feature selection which are discussed comprehensively about attribute selection based on entropy metrics [17, 20, 21]. Some new feature selection methods are derived based on entropy metric such as symmetrical uncertainty, information gain and gain ratio [22] and mutual information [23].

The concept of two dimensional discriminant rule initially grew in association rules generation and also support for classification and regression[7]. The core of the operation is laid on the concept of x-monotone region optimization [4, 7]. The concern of correlation in numeric and discrete class discussed in detail with comparison to accuracy level is considered for single attributes [24].

The need to accommodate pairwise attribute as two dimensional feature is further applied in pattern [25]. Other recent researchers also attempt to explore the pairwise attributes selection in the aspect of noise detection [26].

3. The Performance Framework

Our proposed method filters out the irrelevant attributes to target class by ranking each feature according to the discrimination measure and then select features with high ranking value. The discriminant measure as the separability criterion in the *categorical class* feature selection is provided with utilization of entropy in form of maximization of information gain of discrete target class. The *continuous* attribute selection problem refers to the assignment on numerical attribute with domain threshold range in such a way that the order corresponding to real value is preserved. The separability criterion in continuous target class problem is the minimization of mean square error (MSE), which is equivalent to the maximization of the interclass variance (ICV). Each single metric will be defined in the following descriptions.

Definition 1: In a classification problem, Morimoto et al. [7] defined ***a stamp point*** of pixel grid region R as $\left(\sum_{B_{i,j}\in R} x(B_{i,j}), \sum_{B_{i,j}\in R} y(B_{i,j})\right)$ which are mapped within bucketing operation that takes place for grabbing data item purpose from dataset [5]. We link consecutive bucket B_s, B_{s+1},..,B_t within the data range $[x_s,x_t]$. We assume that we have M records of data, which divided into N buckets almost evenly, the complexity of this procedure requires *O(M log N)*[5]. We assume M be the number of records in the database, for each numeric attributes, we create an *equi-depth* bucketing [2, 3, 8]. Equi-depth bucketing means to generate buckets for each (conditional) attribute. It assures each bucket contains almost same number of data and makes a pixel grid for each pair of attribute. The records are uniformly distributed into $N \le \sqrt{M}$ ordered buckets according to the values of their attributes.

We apply the equi-depth approach to warrant the picking of proportional distribution of data item. Where the discretization into pixel grids follows the rule:

- For classification problem: we count the positive and negative frequency for each pixel as an atomic stamp point, the pair value $(x(B_{i,j}), y(B_{i,j}))$ is for the (i,j) bucket. It is the query of count all attributes occurrence on the positive and negative target class values.
- For regression problem: we count data frequency of data and sum of target class for each pixel as an atomic stamp point, the pair value $(x(B_{i,j}), y(B_{i,j}))$ is for the (i,j) bucket.

Definition 2: Feature selection can be stated as follows. Given an initial set of n features F, F= $\{f_1,f_2,..,f_n\}$, and class label T that we wish to predict from the feature set, find the subset f with δ features where $\delta<n$, that maximize certain performance measure of the prediction performance for a given classifier [1, 6, 7, 10].

Definition 3: In the classification problem which target class is categorical or discrete, the performance goal is to maximize Information Gain (IG) or Relative Entropy. Information gain gives the separability criterion on each pair of attribute and as a single metric for ranking criteria of the optimized regions. This criterion is defined as [5].

$$= \frac{\|X\|}{\|A\|}\sum_{i=1}^{2}\frac{x_i}{\|X\|}\log\frac{x_i}{\|X\|} - \frac{\|A-X\|}{\|A\|}\sum_{i=1}^{2}\frac{a_i-x_i}{\|A-X\|}\log\frac{x_i}{\|A-X\|} \quad (1)$$

where:

$$\|X\| = \sum_{i=1}^{2} x_i; \|A\| = \sum_{i=1}^{2} a_i$$

Where vector $X=(x_1,x_2)$ is a stamp point of a region and vector $A=(a_1,a_2)$ is a stamp point of all data.

Definition 4: In regression problem which target attribute is continuous numeric, the performance goal is to maximize Interclass variance (ICV)[7]. ICV as separability criterion on numeric continuous target class problem is intended as a single metric in attributes ranking process. We confront with stamp point; this ICV is transferable to the new formula as follows

$$ICV = x_1\left(\frac{x_2}{x_1}-\frac{a_2}{a_1}\right)^2 + (a_1-x_1)\left(\frac{a_2-x_2}{a_1-x_1}-\frac{a_2}{a_1}\right)^2 \quad (2)$$

Where vector $X=(x_1,x_2)$ is a stamp point of a region with value of (|R|, sum of target in R) and vector $A=(a_1,a_2)$ is a stamp point of all data. For classification problem, each stamp point represent classification rule whose coordinate value of vector $X=(x_1, x_2)$. Each stamp point is evaluated by Entropy, where $Ent(X) = Ent(x_1, x_2)$. If we have vector $A= (a_1,a_2)$ which is the stamp point of whole data and constant, the optimal region is minimum entropy value which is defined as equation 2. On the other hand for regression problem, each rule is evaluated by ICV, where $ICV(X) = ICV(x_1,x_2)$. Optimal region is the maximal of ICV which is defined as equation 2.

Our experiments are based on two main components. The first one is the idea of x-monotone region optimization, in which we utilize our approach of two dimensional discriminant rule in a form of pairwise attributes optimization using a branch and bound algorithm [4, 5, 7]. The second part is the measurements for classification and regression cases that respectively comprise of maximum information gain and the interclass variance. We provide this framework as a whole to construct the process of defining attribute selection on pairwise attributes.

3.1. X-Monotone Region Optimization

Definition 5: An x-monotone region is a grid region R whose intersection with any vertical line is undivided. Fig.1 gives an x-monotone appearance.

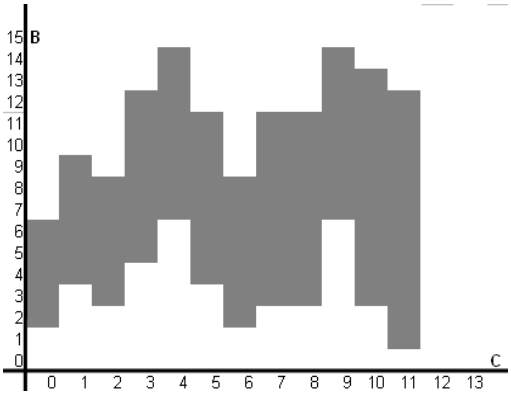

Fig.1 x-monotone region

Definition 6: R_{Opt} is achieved by computing tangent point of the set of stamp points and the line whose normal vector is Θ. Optimal region of x-monotone R will maximize inner product matrix G (Θ, X). The highest index of m-th column containing the s-th row in region that maximizes $\sum_{i=bottom_m(S)}^{s} G_{i,m}$

Definition 7: The $cover_m(s, t)$ on column m is the maximum Σ**G** of stripe on m-th column that containing the s-th and the t-th row, for any s, t, such that 1≤ s, t ≤N

$$\mathrm{cover}_m(s,t) = \begin{cases} \sum_{i=bottom_m(s)}^{top_m(t)} G_{i,m} & s \le 1 \\ \sum_{i=bottom_m(s)}^{top_m(s)} G_{i,m} & s > 1 \end{cases} \quad (3)$$

Definition 8: F_m (j) is the optimal x-monotone region on sub-matrix that maximizes **Σ**G and containing G_{jm}. $F_{m+1}(i)$ is computed from $F_m(1)$, $F_m(2)$…$F_m(N)$ as follows

$$F_{m+1}(i) = \max\{\max(0, F_m(j)) + \mathrm{cover}_m(i,j)\} \quad (4)$$

The following Fig 2 illustrates the $F_m(j)$.

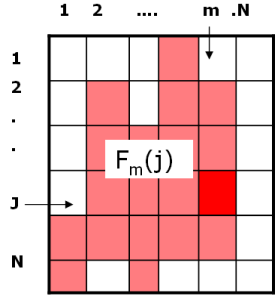

Fig 2. Optimal x-monotone on sub-matrix

Definition 9: The optimal region of a set of stamp point lay on the convex hull, which points are computed efficiently using touching oracle, where any tangent line with the set points touches to a point of convex hull. For any two vector x_1 and x_2 in the domain of vector X, Ent(X) satisfies following inequality[7]:

- $\min\{Ent(x_1), Ent(x_2)\} = Ent((1-\gamma)x1 + \gamma x_2)$
- $\max\{ICV(x_1), ICV(x_2)\} = ICV((1-\gamma)x_1 + \gamma x_2)$ for $0 \le \gamma \le 1$

Assume a line segment in the 2D plane as depict in Fig 2.a. The optimal value on the segment must be on the end of the segment. According to the Fig 2.b. if we consider A,B,C,D,X and Y points, we can say that X<Max(A,B),C < Max(A,Y), Y < Max(B,D) thus only convex Hull points contribute for optimal value.

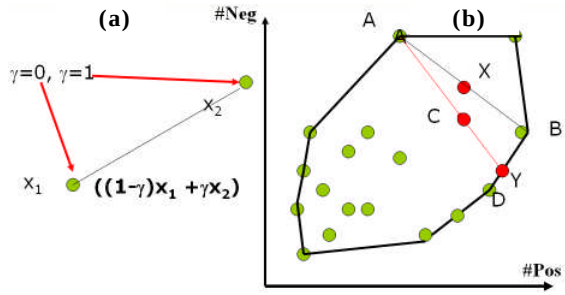

Fig 3. Convex hull properties

In the following Fig 4. We depict the optimization using branch and bound approach by utilize the touching oracle technique in four sector of normal vector Θ.

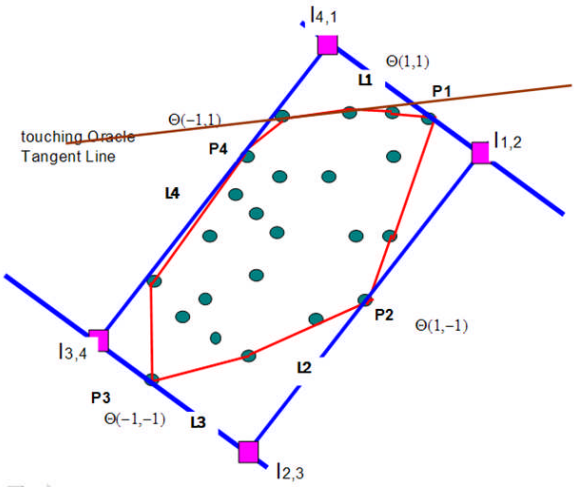

Fig 4. Convex hull with touching oracle

Definition 10: Within all dispersed stamp points, initial four stamp points (P_1,P_2,P_3,P_4) is computed using four vectors θ_1, θ_2, θ_3 and θ_4. The guided Branch and Bound searching [4, 5, 27] is applied on the convex hull using these four vectors to compute four intersecting points of the four tangent lines (L_1,L_2,L_3,L_4) as depicted in Fig 4.

Definition 11: The Entropy (or ICV) values for each intersection ($I_{1,2}$,$I_{2,3}$,$I_{3,4}$,$I_{4,1}$) are computed. The maximum Information Gain (Max(IG)) is the lowest bound of Entropy at each of these values for each sub-chain of the convex hull. (resp. the maximum ICV). This is the best stamp point.

If we construct the procedure of optimized x-monotone region on a N×N grid region, we can express various shapes of data mapping from data set and computed efficiently in range of O(k log k) where $k=N^2$ [6].

3.2. Optimized X-monotone region evaluation

The second core idea in feature selection after separability criterion is selection strategy [25]. A selection strategy in feature selection process conducts further filter upon all optimized x-monotone regions and decide only limited number of pairwise fulfill the requirement as the most informative attributes to the target class.

From this point of view, it means that our selection strategy should be smart to evaluate correlation existence within pairwise attributes according their shape. It is obviously required to compute accurately within acceptable range of values. In order to fulfill the requirement, we briefly define them as follows.

Definition 12: Each optimized x-monotone region shape is evaluated as follows: Sequence pair of vector β and vector τ in N×N 2-D plane buckets. We define two vectors, vector $\tau = \{u_0, u_1,..,u_m\}$ and vector $\beta= \{b_0,b_1,..,b_m\}$;$|\tau|=|\beta|=m$.

The following procedure applies for the region shape of optimized x-monotone region.

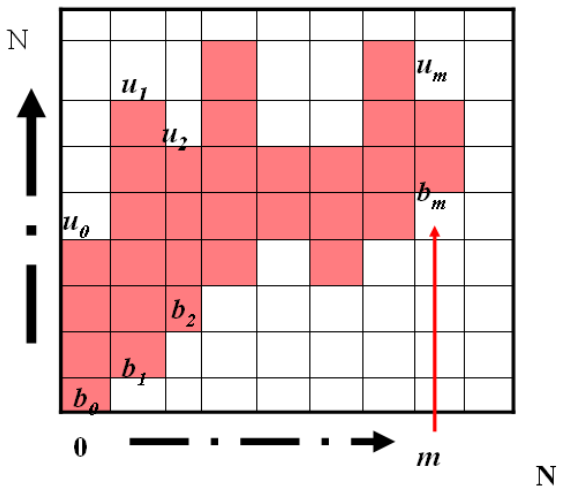

Fig 5. Vector construction of the optimized x-monotone region

Definition 13: In order to evaluate the shape of each optimized region from pairwise attributes, we use the two statistical approximation measurement, skewness and kurtosis metric which constitute with a boundaries set as follows.

Skewness (Skew) is a measure of the asymmetry of the data around the sample mean. Skewness is defined as

$$Skew(\beta) = \frac{m^{-1}\sum_{i=0}^{m}(b_i - \mu)^3}{\sigma^3} \quad (5)$$

Note that similar formula is also applied to the top vector τ of the optimized x-monotone region. μ as means of b_i (respectively u_i), σ as standard deviation of b_i respectively (respectively u_i), m represents the number of vertical stripe of grid.

Empirically we set a tolerance limit as ε which has a value of 1.10^{-4}. We will use this limit as the smallest deviation within shape evaluation of the optimized x-monotone regions, using both skewness and kurtosis.

In order direct the evaluation processes, we set the following boundaries value of skewness as [28]:

- Skew < **-ε** is the data are spread out more to the left (decreasing).
- Skew > **+ε** is the data are spread out more to the right.(increasing)
- **-ε<**Skew**<+ε** when both vector β or τ are tend to form a flat line or equally distributed

Kurtosis (Kurt) is a measure of how outlier-prone a distribution is. Kurtosis is defined as [28, 29]:

$$Kurt(\beta) = \frac{m^{-1}\sum_{i=0}^{m}(\mu - b_i)^4}{\sigma^2} \quad (6)$$

In order to help shape evaluation using kurtosis formulation, the following boundaries applied to the procedure:

- **(3- ε) ≤** Kurt (Normal Distributed) **≤** (3+ **ε)**
- Kurt (More outlier-prone) >(3+ **ε)**;
- Kurt(Less outlier-prone) < (3- **ε)**.

Definition 14: We call an optimized x-monotone region as **trivial**, and as the consequence it is the pairwise attributes of non-correlated features, when the following condition fulfills

- $-\varepsilon \le Skew(\beta) \le +\varepsilon$ or $Skew(\beta) = \infty$ and
- $3.0-\varepsilon \le Kurt(\beta) \le 3.0+\varepsilon$ or $Kurt(\beta = \infty$. Both boundaries also apply into the vector τ.

In the following Fig 6, we depict a common shapes of trivial region.

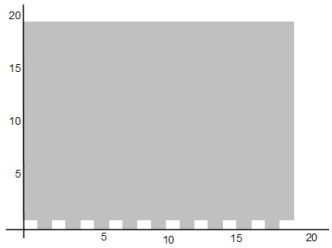

Fig 6. Examples of trivial optimized x-monotone region

Definition 15: A **non trivial** optimized x-monotone region is defined as

- Skew(β)<-ε or Skew(β)> +ε;

And

- Kurt (β) <3 Kurt (β)>3.

As the consequence it is the pairwise attributes of non-correlated features, when the above condition fulfills In the following Fig 7. We provide an example of a non-trivial optimized x-monotone region.

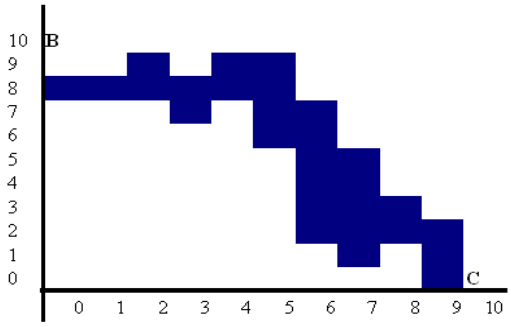

Fig.7 An example of non-trivial optimized x-monotone region

Definition 16: Non-linearity degree (NLD) is our proposed metric that detect the proportional share of non-linearity pairwise attribute within a dataset. The metric is defined as:

$$NDL = \frac{\sum_{x=1}^{R}(\#NonTrivial)}{\sum_{x=1}^{R}(\#Optimized\,\mathrm{Re}gion)} \qquad (7)$$

Number of nontrivial region is counted according to pairwise attribute filter properties from the implementation of our technique to a dataset.

Definition 17: The number of optimized value for categorical problem is defined as $R = \frac{\delta(\delta-1)}{2}$, where R is the number of optimized x-monotone region, reflected from all pairwise attributes in a dataset. The value of x is the number of attributes in a dataset. For continuous problem we set the number of optimized region as $R = \delta(\delta-1)$. As a result, the process of constructing the pairwise in classification and regression require time complexity of $O(\delta^2)$.

4. Pairwise Attributes Selection

In Section 3, we have described the inner component of this algorithm. A remaining issue is how to determine the optimal number of n attributes. Since a mechanism to remove single potentially redundant attributes from the already selected attributes has discussed in previous works [2, 11, 16, 20, 23, 30, 31], we concentrate our work for pairwise attributes which demonstrate a strong correlation between them.

We present a pairwise attributes selection algorithm using the previously developed optimized x-monotone region on two attributes. An algorithm of our key steps is as follows

```
Input  : Dataset in table
Output  :List  Of  Selected  Pairwise
Attributes and Non-Linear Degree of the
Dataset
1.  Generate Stamp Point from N×N Bucketing
2.  Get Optimal Region using Touching Oracle
    of the stamp points
3.  Repeat
    Evaluate Region using Branch and Bound
    Until Max(IG) or Max(ICV)
4.  Do Shape Evaluation using Skew(β) and
    Kurt(β) with ε bound
    If   (-ε≤  Skew(β)  ≤  ε  and  (3-  ε)≤
    Kurt(β)≤(3+ ε))=True Then
    Status=Trivial   Else      Status   =
    NonTrivial
5  While  (Status  =  NonTrivial)Generate
    Attribute Ranking based on Max(IG) or
    Max(ICV)
```

Fig. 8. Key steps of pairwise feature selection

4.1 Selecting Categorical Pairwise Attributes

The core algorithm for selecting the most relevant categorical pairwise attributes consists of initiation, counting all pairs of positive and negative matrix and evaluates all of optimized x-monotone regions results from the pairwise attributes. According algorithm in Fig 8, we called our algorithm as **PAS,** abbreviation of **Pairwise Attributes Selection**. We modify our base algorithm for two purposes; the first one uses the maximization of Information Gain (IG) which solve the correlated attributes on the categorical target class. The second one uses maximization of Interclass Variance (ICV) which solves numerical correlated attributes on

continuous target class.

We construct the guided branch and bound procedure [5, 27] to evaluate candidate region by using the lower bound Entropy calculation in the four intersection points and four vector normal Θ as depicted in Fig 4. These loop end as all stamp points along the convex hull are covered. The results are optimized pair of stamp points that construct an optimized region of x-monotone. The expected runtime is O (k log k).

We follow similar three steps as in categorical problems where each pair of attributes will construct two matrices that are derived from the count of attributes frequency and another matrix from the sum of target class value (definition 2). In this division we generate pairwise attribute ranking using ICV values.

We continue the process with the optimized x-monotone shape evaluation and decide to group all the resulted regions into trivial and non-trivial region as described in Definition 14 and Definition 15. Conventional single attribute selection should not recognize this pairwise attribute as the existence of nonlinear correlation with interdependency will deter their value as unimportant.

We assume that the trivial pairwise attributes as the attributes that sufficiently determinable using conventional method. Their importance to the target class is manageable using single attribute only. Then, we will only focus on the nontrivial results. We generate the pairwise attributes ranking by using their respective value of maximum information or maximum interclass variance.

4.2 Time Complexity Analysis

In this section we analyze the time complexity of PAS. We focus our analysis on the five steps in PAS procedures which are consisted of bucketing procedures, pairwise attribute construction, optimized x-monotone region calculation, shape evaluation and nonlinear degree measurement.

In the bucketing, as defined in **definition 1** we required $O(M \log N)$ in order to dividing M data into N buckets. In the next step of pairwise attributes construction, we require $O(\delta^2)$ to setup all available pair of attributes.

The skimming process for all pairwise attributes using x-monotone region optimization as described in **definition 12,** requires O(k log k) of time complexity, which is continued with shape evaluation using skewness and kurtosis can be computed efficiently on O(Rm), where R is the number of pairwise attributes and m is the number of vertical line with m <N. To compute the overall nonlinear degree in a dataset, requires O(t), where t is the number of nontrivial pair found from optimized x-monotone regions. Thus for all records in a dataset the total complexity of PAS algorithm is $O(M \log N + \delta^2 + k \log k + Rm + t)$.

As a result of the total complexity, we can state that the value of resolution (N) and number of attribute dimension (δ) influence the computational complexity. With this result means lower resolution with lower dimensionality of a dataset reduces the computational cost significantly.

It is interesting that our pairwise attributes selection overall time complexity is a mixed form of logarithmic and linear type. One important point within this overall complexity is that we do not need to do additional effort for optimization as the inner part of x-monotone region have included with optimization using branch and bound algorithm [7]. The following Table 1 gives a brief comparison of complexity cost to other single numeric based feature selection

Table 1. Complexity cost comparison

No	Algorithm and sources	Overall Computational Cost	Description
1	mRMR[32]	O(|S|. Q)	S is # feature set, Q is # feature to select
2	ReliefF[2, 33, 34]	O(M. δ.c)	M is # records, δ is # attributes and c is user's constant
3	Markov Blanket based FS[20]	$O(n^2(M+\log n))$ for computing and sorting attributes and $O(rnmkc2^k)$ for select features	R is number of features to eliminate; c is number of class and k is number of conditioning features
4	Our proposed algorithm	$O(M \log N + \delta^2 + k \log k + Rm + t)$	k is resolution of grid bucket, R is number of pairwise, m is number of vertical line of optimized region and t is number of nontrivial region

The complexity cost comparison above indicates that theoretically our feature selection requires more time to perform whole operation than single attribute selection. The most distinguish point between these algorithms that PAS contain preliminary step of optimization, which ensure the optimal region to gain. Every pair of attributes is warranted to contribute the best value using separability criterion (Max(IG) and Max(ICV) as we utilize the convex hull properties to initiate the process of optimization using branch and bound[7].

5. Implementation and Experiments

In this section we discuss the implementation issues applied on several real dataset. We aim to show that:

1. Our technique can handle and select meaningful features from variety volume of multidimensional real world data sets.
2. Its capability to reveal important features from pairwise attributes approach that missed by conventional methods
3. Any dataset has a portion of strongly correlated attributes that can be very important to the target class determination.

It is essential to highlight that our technique for both categorical and continuous target class were designed to run in multidimensional datasets composed of numeric attributes.

5.1. Dataset up

We tested our attribute selection approach on four discrete and four continuous datasets. We set up them in a MySQL environment in order to ease the data management and retrieval process of bucketing during matrix preparation for further calculation both categorical and continuous target class problems.

The eight datasets we used are shown in Table 2. These datasets are part of UCI machine learning archive. As we refer to recent analog work [25], that pairwise feature evaluation tend to be efficient in a small number of pairwise selection, we compare our proposed method with various datasets with small number of attributes to a mediocre less than 100 attributes. The following Table 2 describes the datasets that we used in this paper:

Table 2 Dataset up for experiment, source UCI ML Repository [35]

No	Datasets	#Attr	#Rec.	#Num Att	Type	Target
1	Pharynx	12	195	12	Con	Class
2	Pollution	15	60	15	Con	Mort
3	Sensory	12	576	11	Con	Score
4	Body fat	15	252	15	Con	Class
5	Adult Census	22	5000	6	Cat	Class
6	Credit	40	40000	7	Cat	Status
7	Pima Diabetes	14	525	8	Cat	Class
8	CoIL200	86	5822	85	Cat	Caravan

Where #Attr indicates the number of attribute within dataset, #Rec is the number of record content, #Num Att is the number of available numeric attribute, the Con in type column is the continuous numeric type and Cat is categorical target class.

As our aim to make a comparison to the current conventional single attribute selection algorithms, we set up similar datasets for attributes selection using two well-known algorithms to rank features of labeled data, Gain ratio (GR) [36] and ReliefF algorithm [2, 33, 34]. For this purpose, we use weka data mining application framework [37]. We ran the experiments using default parameters and 10-fold cross-validation.

Our experiments are performed on a single machine of Pentium-4, 3.00GHz CPU with 1.5GB RAM running the Microsoft Windows XP professional. All datasets are managed by a MySQL5.1 engine and the implementation codes are with Java Standard Edition Release 1.6.0x.

5.2. Experiment Result

5.2.1. Categorical Non-Trivial Result

In the categorical target class problems, we concern on choosing the positive and negative target class label. As we defined, this algorithm will search the entire domain space item within a dataset while building *equi-depth* buckets. The algorithm calculate and determine the most important strongly correlated attributes which not considered on most of conventional feature selection algorithms. The following tables and figure are some results from our experiments.

Table 3. Result of categorical problem of adult census dataset with bucket size of 10

Attributes	Age	Fot	EduNum	CapitaLos	CapitalGain	HourWeek
Age		0.021	0.254	0.608	0.001	0.329
Fot			0.253	0.607	0.000	0.329
EduNum				0.649	0.296	0.465
CapitaLos					0.786	0.678
CapitalGain						0.329
HourWeek						

In the Table 3 above, we notice the pairwise attributes value on Adult census dataset. The figures in the box of the upper triangular are the value of maximal information gain that derived from paired positive and negative matrix of training example. We let the lower triangular cells empty as their values are similar to the upper one as a result of symmetrical properties of convex hull. Non-trivial pairwise attributes denote in white box, whereas the pairwise of (capitalos-edunum) attributes with 0.649 information gain score, is the most important

within this dataset in bucket size of 10. The figures within grey boxes are indicated that these both attributes are similarly important as considered using conventional single attribute selection.

The following Table 4 shows the results from Pima diabetes dataset. Our algorithm succeeded identifying 12 non-trivial pairwise attributes which are important in classifying the target class of diabetes patient. From these 12 pairs we can recognize that the pair of **age** of patient (in year) and 2-hour **serum** insulin is the most important pair within 10 bucket scanning using our PAS algorithm. All of grey boxes figures indicate as non-correlated attributes. For example, patient's **age** with body mass index (**bodydx**) are actually non-correlated attributes, even with higher score max (IG) of 0.749.

Table 4 Result of Pima diabetes in bucket 10

Attributes	Preg	Plas	Dias	Triceps	Serum	BodyIdx	DiabFunc	Age
Preg		0.000	0.140	0.136	0.510	0.078	0.413	0.062
Plas			0.000	0.000	0.000	0.000	0.000	0.000
Dias				0.000	0.000	0.000	0.000	0.000
Triceps					0.560	0.142	0.751	0.169
Serum						0.521	0.719	0.764
BodyIdx							0.725	0.749
DiabFunc								0.749
Age								

Fig.9 shows the result of our computation of the optimized x-monotone region while using credit G dataset. The non-trivial pairwise attributes between client's age and credit shows a nonlinear correlation between these attributes with max (IG) is 0.08 and Skew(β) are 1.4827 for β_i and 1.3443 for τ and then the value of Kurt(β) are 3.8236 for b_i, and 3.3284 for u_i respectively. According to definition 14 and definition 15, we called this region as non-trivial.

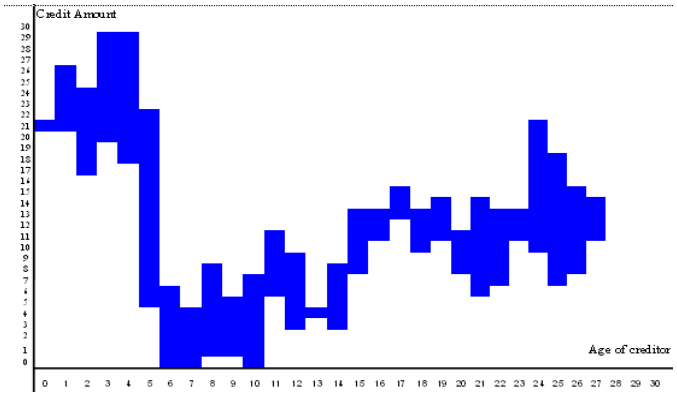

Fig.9 Non-trivial optimized x-monotone region from creditG dataset

5.2.2. Continuous Non-Trivial Result

According to our definition for continuous target class problem, we need to concern about the frequency of attributes and sum of target class value in order to determine the most relevant (resp. must be non-trivial) pairs among a dataset.

The following Fig 10, we provide result of a pairwise attribute which is one of the non-trivial results of important pair between attribute "Age" and "Chest" from body fat dataset in 20 buckets.

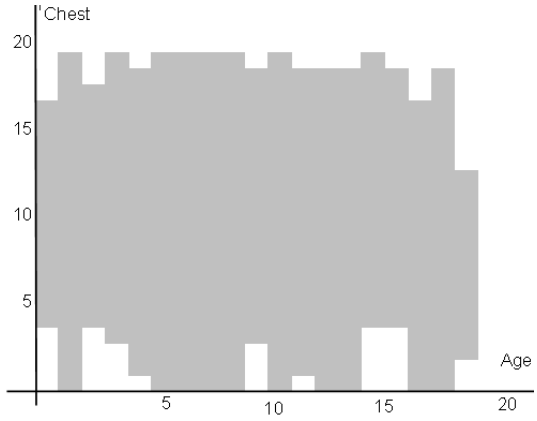

Fig.9 Non-trivial region of continuous class from body fat dataset

Here, the figure shows these two attributes are important to the target class as they show a nonlinear correlation relationship with a higher maximum ICV score. Max (ICV) is 155032.926, Skew (β) are 1.0167 for bi_i and 1.1474 for u_i and then the value of Kurt (β) are 2.6432 for b_i, and 3.2706 for u_i respectively.

6. Discussion

The unique results of our algorithm are the capability to calculate the non-trivial pairwise attributes and show their importance in determining the target class. Furthermore according to our definition 20, currently it is possible for us to detect the non-linearity existence within a dataset. By using the metric of NLD we can measure on what ratio as degree of non-linearity correlation among them. In the following Table 5 shows the ratio as the number of non-trivial pairwise attributes on all optimized region found for each datasets.

According to the distribution and density of data items in the dataset, various numbers of non-trivial pair could be found in different bucket size. For instance in pharynx dataset, within all 132 pairs, we found only 1 pair as nontrivial among 132 optimized x-monotone region generated by PAS in 10 buckets size, then 3 nontrivial pair in 20 buckets, but no pair found in 30 bucket set.

Interesting patterns are also found in other datasets. The urgency of various bucket sizes in this algorithm is to capture the dispersed data items within possible grid size.

Table 5 The Non-Linear Degree results in three buckets set

No	Datasets	# pairwise	Resolution		
			10x	20x	30x
1	Pharynx	132	1/132	3/132	0/55
2	Pollution	210	11/210	9/210	-
3	Sensory	55	2/55	2/55	5/39
4	Body fat	210	41/210	33/210	-
5	Adult Census	15	9/15	-	-
6	Credit G	21	15/21	4/21	3/16
7	Pima Diabetes	28	12/28	6/28	1/7
8	CoIL 2000	7140	20/4450	20/3570	5/59

In this experiment we see that bucket resolution up to 20 or 30 is enough for all datasets we examine. A higher grid resolution means a lower amount of attribute value within each grid. Therefore we expect less to find a non-trivial pairwise attribute.

We can summarize that some datasets are actually non-correlated ones as pharynx, pollution and sensory datasets as their result of lower NLD. But we have to consider carefully with datasets which demonstrate higher NLD ratio such as Body fat and Pima diabetes. These datasets have a strongly correlated and interdependency relationship among their attributes. We can conclude that biological based attributes influence this interdependency as shown from the results.

6.1. Comparison to Other Numeric Algorithms

This section shows how our results are superior to majority of the available conventional attributes selection algorithms. It is interesting to show in what extend these algorithms give the result on the similar dataset. We provide the result of Gain ratio (GR) [36] and ReliefF [2, 33, 34] algorithms for comparison with our PAS on a categorical case of CoIL2000 dataset. In this list we only provide the first-six most relevant attributes on 3 groups' resolution of 10, 20 and 30 in the following Table 6.

Table 6. Comparison of six first attributes selected from the results of conventional feature selection algorithms and *non-trivial* PAS result on CoIL2000 dataset

No	Gain ratio	Relief F	Resolution		
			10	20	30
1	#47	#47	**#1** & #21	**#35**&#57	**#15** & #18
2	#68	#68	**#1** & #64	**#35**&#61	**#15** & #35
3	#61	#59	**#1** & #55	**#35**&#64	**#15** & #39
4	#82	#43	**#1** & #76	**#35**&#57	**#15** & #22
5	#59	#42	**#1** & #70	**#35**&#63	#32 & #62
6	#42	#44	**#1**& #85	#51&#64	#36 & 62

Intuitively, the result of GR should reflect to the result of PAS as their similar basic inner component of entropy value [36]. The GR and ReliefF result the rank of single attributes according to it's respectively score, indicate that attribute #47 is the most important feature to the target class "caravan" in the dataset. Our PAS provides the result with pairwise patterns of two attributes.

According to our definition, we focus on the trivial shape of optimal region only. These two attributes are correlated each other. For example, attribute #1 and # 21 are the most relevant pair to determine the target class of caravan in bucket 10. From bucket 20 our algorithm results with attributes of #35 and #57 then attributes of #15 and #18 in bucket 30. In their respective calculation, attribute #1, #35, #51, #15 and #32 have been scored to zero or not selected by GR and ReliefF methods. Thus, PAS succeeds to reveal the hidden pairs of attributes that actually potential in determine the target class.

According to these results, we observed that our algorithm is able to reveal the hidden pairs of important attributes. Thus we also show that these trivial regions contain nonlinear correlation relationship between attributes in a dataset. Apparently this property has led conventional method to choose as important and relevant attributes.

As these attributes' relationship behavior may occur in many field of domain in real world data transaction, it will be very powerful to reveal hidden important features within the disciplines such as bioinformatics, complex decision making environments and operation research.

7. Conclusion and Future Work

In this paper, we have presented a novel feature selection based on two-dimensional discriminant for finding and selecting a subset of independent and important pairwise attributes on a strongly correlated dataset. Our feature selection could reveal hidden important pairwise attributes that abandoned by conventional algorithms as

their nonlinear correlation properties. These unique results support for both categorical and continuous problems.

In our future work we will further investigate and construct a robust algorithm for x-monotone region pattern approximation to enhance the decision making in related application.

Reference

1. de Sousa, E.P.M., Traina, C., Traina, A.J.M., Wu, L.J., Faloutsos, C.: A fast and effective method to find correlations among attributes in databases. Data Mining and Knowledge Discovery **14** (2007) 367-407
2. Kononenko, I., Hong, S.J.: Attribute selection for modelling. Future Generation Computer Systems **13** (1997) 181-195
3. Liu, W.Z., White, A.P.: The Importance of Attribute Selection Measures in Decision Tree Induction. Machine Learning **15** (1994) 25-41
4. Fukuda, T., Morimoto, Y., Morishita, S., Tokuyama, T.: Data mining with optimized two-dimensional association rules. Acm Transactions on Database Systems **26** (2001) 179-213
5. Fukuda, T., Morimoto, Y., Morishita, S., Tokuyama, T.: Mining optimized association rules for numeric attributes. Journal of Computer and System Sciences **58** (1999) 1-12
6. Fukuda, T., Morimoto, Y., Morishita, S., Tokuyama, T.: Interval finding and its application to data mining. Algorithms and Computation **1178** (1996) 55-64
7. Morimoto, Y., Ishii, H., Morishita, S.: Efficient construction of regression trees with range and region splitting. Machine Learning **45** (2001) 235-259
8. Molina, L.C., Belanche, L., Nebot, A.: Attribute Selection Algorithms: A survey and experimental evaluation. Proceedings of 2nd IEEE's KDD **2002** (2002) 306-313
9. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research **3** (2003) 1157-1182
10. Liu, H.A., Setiono, R.: Incremental feature selection. Applied Intelligence **9** (1998) 217-230
11. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97** (1997) 273-324
12. Bhavani, S.D., Rani, T.S., Bapi, R.S.: Feature selection using correlation fractal dimension: Issues and applications in binary classification problems. Applied Soft Computing **8** (2008) 555-563
13. Haindl, M., Somol, P., Ververidis, D., Kotropoulos, C.: Feature selection based on mutual correlation. Progress in Pattern Recognition, Image Analysis and Applications, Proceedings **4225** (2006) 569-577
14. Liu, H., Motoda, H., Yu, L.: A selective sampling approach to active feature selection. Artificial Intelligence **159** (2004) 49-74
15. Liu, H., Yu, L., Dash, M., Motoda, H.: Active feature selection using classes. Advances in Knowledge Discovery and Data Mining **2637** (2003) 474-485
16. Yu, L., Liu, H.: Feature Selection for High-Dimensional Data: A Fast Correlation-based Filter Solution. Proc. Int.Conference ICML2003 **2003** (2003) 856-863
17. Dash, M., Liu, H.: Feature selection for classification. Intelligent Data Analysis: An International Journal **1** (1997) 131-156
18. Park, J.S., Shazzad, K.M., Kim, D.S.: Toward modeling lightweight intrusion detection system through correlation-based hybrid feature selection. Information Security and Cryptology, Proceedings **3822** (2005) 279-289
19. Dash, M., Liu, H., Motoda, H.: Consistency based feature selection. Knowledge Discovery and Data Mining, Proceedings **1805** (2000) 98-109
20. Koller, D., Sahami, M.: Toward Optimal Feature Selection. Proc. Int.Conference ICML'96 (1996) 170-178
21. Bakus, J., Kamel, M.S.: Higher order feature selection for text classification. Knowledge and Information Systems **9** (2006) 468-491
22. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. Ieee Transactions on Knowledge and Data Engineering **17** (2005) 491-502
23. Peng, H.C., Long, F.H., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. Ieee Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1226-1238
24. Hall, M.A.: Correlation based Feature Selection for Discrete and Numeric class Machine Learning. In Proceedings of 17th International Conf. on Mach.Learn. (2000)
25. Harol, A., Lai, C., Pezkalska, E., Duin, R.P.W.: Pairwise feature evaluation for constructing reduced representations. Pattern Analysis and Applications **10**

(2007) 55-68
26. Van Hulse, J.D., Khoshgoftaar, T.M., Huang, H.Y.: The pairwise attribute noise detection algorithm. Knowledge and Information Systems **11** (2007) 171-190
27. Fukuda, T., Morimoto, Y., Morishita, S., Tokuyama, T.: Interval finding and its application to data mining. IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences **E80a** (1997) 620-626
28. Joro, T., Na, P.: Portfolio performance evaluation in a mean-variance-skewness framework. European Journal of Operational Research **175** (2006) 446-461
29. Velasco, F., Verma, S.P.: Importance of skewness and kurtosis statistical tests for outlier detection and elimination in evaluation of geochemical reference materials. Mathematical Geology **30** (1998) 109-128
30. Liu, H., Setiono, R.: Feature selection via discretization. Ieee Transactions on Knowledge and Data Engineering **9** (1997) 642-645
31. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research **5** (2004) 1205-1224
32. Peng, H.C., Ding, C., Long, F.H.: Minimum redundancy - Maximum relevance feature selection. Ieee Intelligent Systems **20** (2005) 70-71
33. Bratko, I., Cestnik, B., Kononenko, I.: Attribute-based learning. Ai Communications **9** (1996) 27-32
34. Robnik-Sikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning **53** (2003) 23-69
35. UCI-Learning-Repository: UCI Learning Repository. (2005)
36. Quinlan, J.R.: C4.5 Programs for Machine Learning. (1993)
37. Witten, I.H., Frank, E.: Data Mining-Practical Machine Learning Tools and Techniques with Java Implementation 2nd Edition. (2004)