Minimum Variance Associations — Discovering Relationships in Numerical Data

Szymon Jaroszewicz

National Institute of Telecommunications
Warsaw, Poland
s.jaroszewicz@itl.waw.pl

Abstract. The paper presents minimum variance patterns: a new class of itemsets and rules for numerical data, which capture arbitrary continuous relationships between numerical attributes without the need for discretization. The approach is based on finding polynomials over sets of attributes whose variance, in a given dataset, is close to zero. Sets of attributes for which such functions exist are considered interesting. Further, two types of rules are introduced, which help extract understandable relationships from such itemsets. Efficient algorithms for mining minimum variance patterns are presented and verified experimentally.

1 Introduction and Related Research

Mining association patterns has a long tradition in data-mining. Most methods, however, are designed for binary or categorical attributes. The usual approach to numerical data is discretization [22]. Discretization however leads to information loss and problems such as rules being split over several intervals. Approaches allowing numerical attributes in rule consequent have been proposed, such as [3,25], but they do not allow undiscretized numerical attributes in rule antecedent.

Recently, progress has been reported in this area, with a number of papers presenting extensions of the definition of support not requiring discretization [23,14,7]. Other papers provide alternative approaches which also do not require discretization [20,12,19,1,5].

This work extends those methods further, allowing for the discovery of complex nonlinear relationships between sets of numerical attributes without the need for discretization. The work is set in the spirit of association rule mining. First, a concept of minimum variance itemsets is introduced. Those itemsets describe functions which are always close to zero on a given dataset, and thus represent equations describing relationships in data. Based on those itemsets, rules can be derived showing relationships between disjoint sets of attributes. An Apriori style mining algorithm is also presented.

Let us now review the related work. The approach presented in [16] allows for combining attributes using arithmetic operations, but after combining them discretization is applied. Also, since only addition and subtraction are allowed, nonlinear relationships cannot be represented.

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 172–183, 2008.

In [20,12,19] a method for finding rules of the form "if a linear combination of some attributes is above a given threshold, then a linear combination of another set of attributes is above some other threshold" is described. Rules of this type are mined using standard optimization algorithms. While the approach could be extended to nonlinear case, the method presented here is more efficient since it requires solving eigenvalue problems of limited size instead of using general optimization methods on the full dataset. Furthermore, since binary thresholds are used, the method from [20] cannot represent continuous relationships between groups of attributes. Our work is more in the standard association rule spirit providing both itemsets and rules, as well as an Apriori style mining algorithm.

In [1], an interesting method is presented for deriving equations describing clusters of numerical data. The authors first use a clustering algorithm to find correlation clusters in data, and then derive equations describing the linear space approximating each cluster's data points based on the cluster's principal components computed using eigenvectors of the correlation matrix of data in the cluster. While the use of eigenvectors to discover equations may suggest similarities, the approach presented here is quite different. We are not trying to describe previously discovered clusters, but give method of pattern discovery (defining itemsets and rules) in the spirit of association rule mining. Further we allow for arbitrarily complex nonlinear relationships to be discovered, while [1] essentially describes a cluster as a linear subspace. Third, by adding an extra constraint to the optimization, we guarantee that patterns discovered will not involve statistically independent attributes.

There is some similarity between our approach and equation discovery [9,18]. Equation discovery algorithms are in principle capable of discovering minimum variance patterns we propose. However the discovery methodology, is quite different in both cases. In fact our approach was more than an order of magnitude more efficient than Lagrange [9], an equation discovery system. Combining the two approaches, such as using equation discovery to give explicit formulas for minimum variance patterns is an interesting topic for future research.

2 Minimum Variance Itemsets

Let us begin by introducing the notation and some preliminary concepts.

We assume that we are dealing with a dataset D whose attributes are all numeric. Non-numerical attributes can be trivially converted to $\{0, 1\}$ attributes. To avoid overflow problems while computing powers, we also assume that the attributes are scaled to the range $[-1, 1]$.

Attributes of D will be denoted with letters X with appropriate subscripts, and sets of attributes with letters I, J, K. If $t \in D$ is a record of D, let $t.X$ denote the value of attribute X in t, and $t[I]$ the projection of t on a set of attributes I. Following [15,8] we now define support of arbitrary functions. Let f be a function of an attribute set I. Support of f in D is defined as

$$\mathrm{supp}_D(f) = \sum_{t \in D} f(t[I]).$$

We are now ready to describe minimum variance itemsets, the key concept of this work. Our goal is to discover arbitrary relationships between the attributes of D. The patterns we are looking for have the general form

$$f(I) = f(X_1, X_2, \ldots, X_r),$$

where we expect the function f to somehow capture the relationship among the variables of $I = \{X_1, X_2, \ldots, X_r\}$.

Let us look at two examples. Suppose we have two attributes x and y, such that $x = y$. The equality between them can be represented by an equation

$$f(x, y) = x - y = 0,$$

so one possible function f for this case is $x - y$. Suppose now that x, y represent random points on a circle of radius 1. The function f could now be $f(x, y) = x^2 + y^2 - 1$ since the relationship can be described by an equation $x^2 + y^2 - 1 = 0$. Of course if noise was present the equalities would be satisfied only approximately.

The common pattern of the two above cases is, that the function f was identically equal to zero for all points (records) in the data. It is thus natural, for a given itemset I, to look for a function $f(I)$ which minimizes

$$\sum_{t \in D} [f(t[I])]^2 = \mathrm{supp}_D(f^2).$$

We will call this quantity the *variance of f around zero*, or briefly *variance*, and a function minimizing it, a *minimum variance itemset*. This concept should not be confused with statistical notion of variance, which would be around the function's mean (we consciously abuse the terminology).

This formulation has a problem. The function $f(I) \equiv 0$ minimizes variance but does not carry any information. Also $\frac{1}{2}f$ necessarily has lower variance than f, although it does not carry any more information. To avoid such situations, we add a normalizing condition guaranteeing that the function f is of appropriate magnitude. Several such normalizations will be presented below.

2.1 Formal Problem Statement

The above discussion was in terms of arbitrary functions. In practice we have to restrict the family of functions considered. Here we choose to approximate the functions using polynomials, such that the degree of every variable does not exceed a predefined value d. Let $I = \{X_1, \ldots, X_r\}$ be a set of attributes. Then any function f of interest to us can be represented by

$$f_{\mathbf{c}}(I) = f(X_1, \ldots, X_r) = \sum_{\alpha_1=0}^{d} \cdots \sum_{\alpha_r=0}^{d} c_{(\alpha_1,\ldots,\alpha_r)} X_1^{\alpha_1} \cdots X_r^{\alpha_r},$$

where $c_{(\alpha_1,\ldots,\alpha_r)}$ are the coefficients of the polynomial. We will organize all coefficients and monomials involved in two column vectors (using the lexicographic ordering of exponents):

$$\mathbf{c} = [c_{(0,\ldots,0)}, c_{(0,\ldots,1)}, \ldots, c_{(d,\ldots,d)}]^T,$$
$$\mathbf{x} = [X_1^0 \cdots X_r^0,\ X_1^0 \cdots X_r^1,\ \ldots,\ X_1^d \cdots X_r^d]^T.$$

We now have $f_{\mathbf{c}} = \mathbf{c}^T\mathbf{x} = \mathbf{x}^T\mathbf{c}$, and ${f_{\mathbf{c}}}^2 = \mathbf{c}^T(\mathbf{x}\mathbf{x}^T)\mathbf{c}$. Notice that $\mathbf{x}\mathbf{x}^T$ is a $(d+1)^r \times (d+1)^r$ matrix, whose entries are monomials with each variable raised to power at most $2d$. So the entry in row corresponding to $(\alpha_1, \ldots, \alpha_r)$ and column corresponding to $(\beta_1, \ldots, \beta_r)$ is $X_1^{\alpha_1+\beta_1} \cdots X_r^{\alpha_r+\beta_r}$.

We now use the trick from [15] in order to compute support of ${f_{\mathbf{c}}}^2$ for various values of $\mathbf{c}$ without accessing the data. Let $t[\mathbf{x}]$ denote the $\mathbf{x}$ vector for a given record t, *i.e.* $t[\mathbf{x}] = [t.X_1^0 \cdots t.X_r^0,\ t.X_1^0 \cdots t.X_r^1,\ \ldots,\ t.X_1^d \cdots t.X_r^d]^T$. Now

$$\mathrm{supp}_D({f_{\mathbf{c}}}^2) = \sum_{t \in D} \mathbf{c}^T(t.\mathbf{x} \cdot t.\mathbf{x}^T)\mathbf{c} = \mathbf{c}^T \left(\sum_{t \in D} t.\mathbf{x} \cdot t.\mathbf{x}^T \right) \mathbf{c} = \mathbf{c}^T \mathbf{S}_D\, \mathbf{c}, \qquad (1)$$

where $\mathbf{S}_D$ is a $(d+1)^r \times (d+1)^r$ matrix, whose entry in row corresponding to $(\alpha_1, \ldots, \alpha_r)$ and column corresponding to $(\beta_1, \ldots, \beta_r)$ contains the value of $\mathrm{supp}_D(X_1^{\alpha_1+\beta_1} \cdots X_r^{\alpha_r+\beta_r})$. It thus suffices to compute supports of all necessary monomials, after which support of ${f_{\mathbf{c}}}^2$ for any coefficient vector $\mathbf{c}$ can be computed without accessing the data, using the quadratic form (1).

We now go back to the problem of normalizing $f_{\mathbf{c}}$ such that the trivial solution $f_{\mathbf{c}} \equiv 0$ is avoided. We tried various normalizations:

(a) require that the vector $\mathbf{c}$ be of unit length, $||\mathbf{c}|| = 1$,
(b) require that weighted length of $\mathbf{c}$ be 1, $\sum_\alpha w_\alpha {c_\alpha}^2 = 1$, this allows for penalizing high degree coefficients.
(c) require that support of ${f_{\mathbf{c}}}^2(I)$ be equal to one, under the assumption that all variables in I are distributed uniformly.
(d) require that support of ${f_{\mathbf{c}}}^2(I)$ be equal to one, under the assumption that all variables in I are distributed as in D, but are independent.

When no outliers were present, all of those approaches worked reasonably well. However in the presence of outliers only approach (d) was useful. Other methods picked $f_{\mathbf{c}}$ such that it was close to zero everywhere except for the few outlier points. Also, this approach guarantees that patterns involving statistically independent attributes will have high minimum variance.

We thus limit further discussion to normalization based on the requirement (d). Imagine a hypothetical database D_I in which each attribute is distributed as in D but all attributes are independent. The support of ${f_{\mathbf{c}}}^2$ under such an independence assumption can be computed analogously to (1) as $\mathrm{supp}_I({f_{\mathbf{c}}}^2) = \mathbf{c}^T \mathbf{S}_I\, \mathbf{c}$, where an element of $\mathbf{S}_I$ in row corresponding to $(\alpha_1, \ldots, \alpha_r)$ and column corresponding to $(\beta_1, \ldots, \beta_r)$ is given by

$$\mathrm{supp}_{D_I}(X_1^{\alpha_1+\beta_1} \cdots X_r^{\alpha_r+\beta_r}) = \mathrm{supp}_D(X_1^{\alpha_1+\beta_1}) \cdots \mathrm{supp}_D(X_r^{\alpha_r+\beta_r}),$$

since variables $X_1, \ldots, X_r$ are assumed to be independent.

We are now ready to formally define a minimum variance itemset for a given set attributes I:

Definition 1. *A real valued function f on a set of attributes I is called* itemset *on I. The* variance *of f is defined as*

$$\mathrm{var}(f) = \mathrm{supp}_D(f^2).$$

A minimum variance itemset *on I is a function $f^*(I) = f_{\mathbf{c}^*}(I)$ on a set of attributes I which minimizes $\mathbf{c}^T \mathbf{S}_D \mathbf{c}$ subject to a constraint $\mathbf{c}^T \mathbf{S}_I \mathbf{c} = 1$.*

2.2 Finding the Minimum Variance Itemset for a Set of Attributes

To find a minimum variance itemset for a given I we use the method of Lagrange multipliers [11]. The Lagrangian is $L(\mathbf{c}, \lambda) = \mathbf{c}^T \mathbf{S}_D \mathbf{c} - \lambda \left(\mathbf{c}^T \mathbf{S}_I \mathbf{c} - 1\right)$. Using elementary matrix differential calculus [24,13] we get $\frac{\partial L}{\partial \mathbf{c}} = 2\mathbf{S}_D \mathbf{c} - 2\lambda \mathbf{S}_I \mathbf{c}$, and after equating to zero we get the necessary condition for the minimum:

$$\mathbf{S}_D \mathbf{c} = \lambda \mathbf{S}_I \mathbf{c}. \tag{2}$$

This is the generalized eigenvalue problem [10,24,13], well studied in computational linear algebra. Routines for solving this problem are available for example in LAPACK [10]. If $(\mathbf{c}, \lambda)$ is a solution to (2), a candidate solution $\mathbf{c}'$ to our optimization problem is obtained by scaling $\mathbf{c}$ to satisfy the optimization constraint: $\mathbf{c}' = \frac{\mathbf{c}}{\sqrt{\mathbf{c}^T \mathbf{S}_I \mathbf{c}}}$. Variance of this solution (using substitution and Equation 2) is

$$\mathrm{var}(f_{\mathbf{c}'}) = \mathrm{supp}_D({f_{\mathbf{c}'}}^2) = {\mathbf{c}'}^T \mathbf{S}_D \mathbf{c}' = \frac{\mathbf{c}^T}{\sqrt{\mathbf{c}^T \mathbf{S}_I \mathbf{c}}} \cdot \frac{\mathbf{S}_D \mathbf{c}}{\sqrt{\mathbf{c}^T \mathbf{S}_I \mathbf{c}}} = \frac{\lambda \mathbf{c}^T \mathbf{S}_I \mathbf{c}}{\mathbf{c}^T \mathbf{S}_I \mathbf{c}} = \lambda.$$

The variance of $\mathbf{c}'$ is thus equal to the corresponding eigenvalue, so the final solution $\mathbf{c}^*$ is the (scaled) eigenvector corresponding to the smallest eigenvalue.

The above property can be used to speed up computations, since finding only the smallest eigenvalue can be done faster than finding all eigenvalues (routines for finding a subset of eigenvalues are also available in LAPACK).

Another important observation is that matrices S_D and S_I are symmetric (follows directly from their definition) and positive semi-definite (support of a square of a function cannot be negative). This again allows for more efficient computations, see [10,24] for details.

2.3 Example Calculation

We will now show an example calculation on a toy example of a dataset $D = \{(1,-2),(-2,4),(-1,2)\}$ over attributes x, y, for $d = 1$. $\mathbf{x} = [1, x, y, xy]^T$, and $\mathbf{c} = [c_{(0,0)}, c_{(1,0)}, c_{(0,1)}, c_{(1,1)}]^T$. Now, $\mathrm{supp}_D(1) = 3, \mathrm{supp}_D(x) = -2, \mathrm{supp}_D(y) = 4$, $\mathrm{supp}_D(xy) = -12$, $\mathrm{supp}_D(x^2) = 6$, $\mathrm{supp}_D(y^2) = 24$, $\mathrm{supp}_D(x^2y) = 16$, $\mathrm{supp}_D(xy^2) = -32$, $\mathrm{supp}_D(x^2y^2) = 72$. Supports under independence assumption are $supp_I(y) = \mathrm{supp}_D(x^0) \cdot \mathrm{supp}_D(y) = 12$, $\mathrm{supp}_I(x^2y) = \mathrm{supp}_D(x^2) \cdot \mathrm{supp}_D(y) = 24$, etc. The $\mathbf{S}_D$ and $\mathbf{S}_I$ matrices are

$$\mathbf{S}_D = \begin{bmatrix} 3 & -2 & 4 & -12 \\ -2 & 6 & -12 & 16 \\ 4 & -12 & 24 & -32 \\ -12 & 16 & -32 & 72 \end{bmatrix}, \quad \mathbf{S}_I = \begin{bmatrix} 9 & -6 & 12 & -8 \\ -6 & 6 & -8 & 24 \\ 12 & -8 & 24 & -48 \\ -8 & 24 & -48 & 144 \end{bmatrix}.$$

After solving the generalized eigenvalue problem and rescaling we get $\mathbf{c}^* = [0, -0.5, -0.25, 0]$. The correct relationship $-2x - y = 0$ has been discovered.

Let us now discuss closure properties of minimum variance itemsets.

Theorem 1. *Let $I \subseteq J$ be two sets of attributes, and $f^*(I)$ and $g^*(J)$ be minimum variance itemsets on I and J respectively. Then* $\mathrm{var}(g^*) \leq \mathrm{var}(f^*)$.

In other words variance is upward closed, adding attributes reduces the variance. The proof is a trivial consequence of the fact that a function of I is also a function of J (constant in variables in $J \setminus I$), so the lowest variance attainable for J is at least as low as the variance attainable for I, and may be better.

The problem is that we are interested in itemsets with low variance, so if one is found, all its supersets are potentially interesting too. The solution is to set a minimum threshold for variance, and then find smallest (in the sense of set inclusion) sets of attributes for which the variance (of the minimum variance itemset or the itemset's best equality or regression rule) is less than the specified threshold. Similar approach has been used *e.g.* in [6]. The algorithm is a simple adaptation of the Apriori algorithm [2], and is omitted due to lack of space.

3 From Itemsets to Rules

In order to facilitate the interpretation of minimum variance itemsets two types of rules are introduced. The first kind are what we call *equality rules*.

Definition 2. *An* equality rule *is an expression of the form $g(I) = h(J)$, where $I \cap J = \emptyset$, and g and h are real valued functions on I and J respectively. The* variance *of the rule is defined as* $\mathrm{var}(g(I) = h(J)) = \mathrm{supp}_D\left((g-h)^2\right)$.

Thus equality rules capture relationships between disjoint groups of attributes which are usually easier to understand than the itemsets defined above.

A minimum variance equality rule $g^*(I) = h^*(J)$ is defined, similarly to the minimum variance itemset case above, as a pair of functions for which $\mathrm{var}\left(g^*(I) = h^*(J)\right)$ is minimum subject to a constraint that the support of $(g-h)^2$ is equal to one, under the independence assumption. Finding minimum variance equality rules for given I and J can be achieved using the same approach as finding minimum variance itemsets. If we approximate both g and h with polynomials, $I = \{X_1, \ldots, X_r\}$ and $J = \{X_{r+1}, \ldots, X_{r+s}\}$, and denote

$$
\begin{aligned}
\mathbf{c}_g &= [c_{(0,\ldots,0)}, c_{(0,\ldots,1)}, \ldots, c_{(d,\ldots,d)}]^T, \\
\mathbf{x}_g &= [X_1^0 \cdots X_r^0, X_1^0 \cdots X_r^1, \ldots, X_1^d \cdots X_r^d]^T, \\
\mathbf{c}_h &= [d_{(0,\ldots,1)}, d_{(0,\ldots,2)}, \ldots, d_{(d,\ldots,d)}]^T, \\
\mathbf{x}_h &= [X_{r+1}^0 \cdots X_{r+s}^1, X_{r+1}^0 \cdots X_{r+s}^2, \ldots, X_{r+1}^d \cdots X_{r+s}^d]^T,
\end{aligned}
$$

we get $g = \mathbf{c}_g^T \mathbf{x}_g$, $h = \mathbf{c}_h^T \mathbf{x}_h$, and $g + h = [\mathbf{c}_g | \mathbf{c}_h]^T \cdot [\mathbf{x}_g | \mathbf{x}_h]$. Note that the constant term is omitted from $\mathbf{c}_h$ and $\mathbf{x}_h$, since it is included in $\mathbf{c}_g$ and $\mathbf{x}_g$.

From that point on, the derivation proceeds exactly as in the case of minimum variance itemsets in order to find the vector $[\mathbf{c}_g|\mathbf{c}_h]^*$ which minimizes $\text{supp}_D\left((g+h)^2\right)$ subject to $\text{supp}_I\left((g+h)^2\right) = 1$. After finding the solution, signs of coefficients in $\mathbf{c}_h$ are reversed to get from a minimum variance for $g+h$ to the desired minimum variance for $g-h$.

Finding a minimum variance equality rule on I and J is analogous to finding a minimum variance itemset f on $I \cup J$ subject to an additional constraint that f be a difference of functions on I and J. Thus, the minimum variance of an itemset on $I \cup J$ is less than or equal to the minimum variance of an equality rule on I and J. If an itemset has high minimum variance, we don't need to check rules which can be generated from it, since their variance is necessarily high too.

Another kind of rules are what we call *regression rules*.

Definition 3. *A* regression rule *is an expression of the form $X = g(I)$, where X is an attribute, I a set of attributes, $X \notin I$, and g is a function of I.*

It is easy to see that regression rules are equality rules with additional constraint that one side of the rule must contain a single attribute in the first power only. It is thus clear that minimum variance of a regression rule cannot be lower than minimum variance of a corresponding equality rule. Also, the definition of variance of a regression rule as well as discovery of minimum variance regression rules are analogous to the case of equality rules and are thus omitted.

Minimum variance regression rules correspond to standard least-squares polynomial regression with X being the dependent variable. Therefore minimum variance equality rules can be seen as a generalization of standard polynomial regression to allow functions of dependent variables, and minimum variance itemset as a further generalization allowing for discovering patterns not involving equality.

4 Illustrative Examples

In this section we show some illustrative examples of patterns discovered, and give some suggestions on how to elicit understandable knowledge from them.

We first apply the method to a small artificial dataset. The dataset has three attributes x, y, z, and is generated as follows: (x, y) are randomly chosen points on a unit circle and z is set equal to x. The relationships among the attributes are therefore $z = x$, $x^2 + y^2 = 1$, and $z^2 + y^2 = 1$.

We applied the algorithm with $d = 2$ without any minimum variance threshold. Only pairs of attributes were considered. Generated patterns are given in the table below (terms with negligibly small coefficients are omitted)

attrs.	min. variance	equation
$\{x, y\}$	$6.62 \cdot 10^{-15}$	$-1.99 + 1.99x^2 + 1.99y^2$
$\{y, z\}$	$6.62 \cdot 10^{-15}$	$-1.99 + 1.99y^2 + 1.99z^2$
$\{x, z\}$	$1.24 \cdot 10^{-17}$	$-0.663x^2 + 1.325xz - 0.663z^2 = -0.663(x - z)^2$

The minimum variance itemsets for $\{x, y\}$ and $\{y, z\}$ do not require any comment. They clearly capture the correct relationship $x^2 + y^2 = 1$.

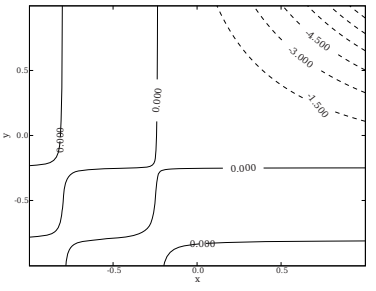

Fig. 1. Contour of the minimum variance itemset for random points satisfying the condition $x < 0 \vee y < 0$

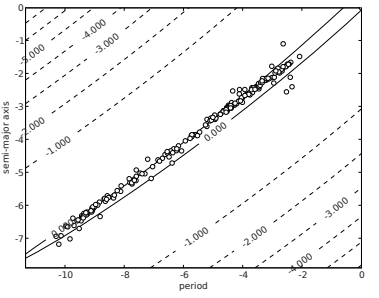

Fig. 2. Extrasolar planet data: relationship between logarithms of planet's period and semi-major axis, together with contours of the minimum variance itemset. Solid line is the zero contour.

The case for $\{x, z\}$ is more interesting. Instead of the expected $x - z = 0$ we obtained an equivalent, but more complicated expression $(x - z)^2 = 0$. The reason is that the degree of the approximating polynomial exceeds that of the true relationship. As a result, two of the eigenvalues are equal to zero, and any linear combination of their corresponding eigenvectors is also a minimum variance solution. To avoid such situations we recommend decreasing the value of d until a minimum value is found at which the relationship still occurs. In the currently analyzed case lowering d to 1 gives the expected $-0.997x + 0.997z = 0$. Another approach, to use regression rules, which also helps in this case.

It should be noted that the best regression rules for $\{x, y\}$ and $\{y, z\}$ have variance of about 1, so the relationship would not have been discovered by standard regression analysis (indeed the correlation coefficient is about $8 \cdot 10^{-3}$).

Let us look at another example which shows that minimum variance itemsets are able to represent patterns much reacher than those usually described using algebraic equations. Consider an artificial dataset which has two attributes $x, y \in [-1, 1]$ and contains points randomly generated on the set where the condition $x < 0 \vee y < 0$ is true. Thus no points are present in the $[0, 1] \times [0, 1]$ square. The correlation coefficient is -0.359, thus not very high. The minimum variance itemset on xy however, has small values everywhere except for the $[0, 1] \times [0, 1]$ square and the minimum variance of $\{x, y\}$ is 0.024. The representation is of course not perfect, but tends to approximate the data quite well (Figure 1). We will see a similar pattern occurring in real life datasets (`sonar`) below.

Extrasolar planets data. This section shows more examples of minimum variance patterns. The dataset used is about currently known extrasolar planets, and can be downloaded from [21]. Six attributes were chosen and 197 planets selected for which all those attributes were defined. The attributes are described in the table below:

attribute	description
pl. mass	mass of the planet
period	orbital period around star
semi-major axis	distance of the planet from star
ang. distance	angular distance of planet from star (as seen from Earth)
star distance	distance of planet's star from Earth
star mass	mass of the star

Attributes were scaled to $[0, 1]$ range, so units are omitted. Afterwards, logarithmic transform was applied. The advantage of the data is that there are some well established relationships which should be discovered if the method works correctly. This experiment is similar to that from [18], but uses more data and involves additional relationships.

First, semi-major axis divided by the distance of the star from Earth is equal to the tangent of the angular distance of the star from the planet. Second, by Kepler's law, the square of orbital period of a planet is proportional to the cube of the semi-major axis of its orbit. If planet and star masses are known, the proportionality constant can also be determined [17]. It is possible that further relationships exist, but due to the author's lack of astronomical knowledge they will not be discussed. We begin by looking at pairs of attributes. The value $d = 2$ was used, with no minimum variance requirement.

The strongest relationship was discovered between planet's `period` and its `semi-major axis` with minimum variance of $6.83 \cdot 10^{-5}$. The relationship is shown in Figure 2. The data points are marked with circles. Contour plot of the minimum variance itemset is also shown. According to Kepler's law there is a linear relationship between logarithms of the two values. The minimum variance itemset is not linear (due to overfitting and ignoring the star mass) but captures the relationship well. Decreasing the degree or examining rules, reveals the linear nature. The clarity of the relationship is surprising since, planet and star masses also play a role. It turned out, that masses of most stars in the data are very close to each other, and planets' masses are too small to distort the relationship.

To explore the relationship further we examined patterns of size 3 and 4 containing attributes `period` and `semi-major axis`. As expected, the most interesting pattern of length three added the `star mass` attribute (minimum variance $6.85 \cdot 10^{-7}$), and by adding `pl. mass`, a four attribute set was obtained with variance $8.33 \cdot 10^{-10}$ — an almost perfect match.

The triple of attributes which had the lowest variance of $9.72 \cdot 10^{-8}$ was `semi-major axis`, `ang. distance` and `star distance`. This is expected due to the deterministic relationship among them described above. All equality rules involving those attributes had very low variance too. Variance of regression rules was higher (in the range of 10^{-4}).

An interesting subset of the above triple, is the pair `semi-major axis` and `ang. distance`. Its minimum variance is 0.027, but the variance of all rules between those attributes is much higher, about 0.15 in all cases. This is another example of a low variance itemset which cannot be captured by equality rules. The situation is depicted graphically in Figure 3, where data points and the

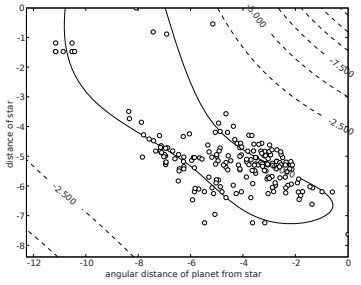

Fig. 3. Relationship between `semi-major axis` and `ang. distance` for the extrasolar planet data, and the corresponding minimum variance itemset. Solid line is the zero contour.

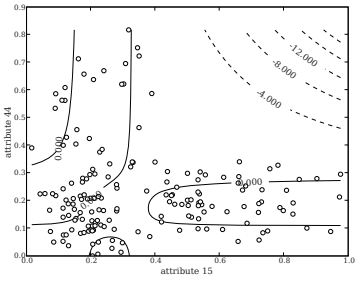

Fig. 4. Data point and contour of the minimum variance itemset for attributes 15 and 44 of the `sonar` dataset

contours of the itemset are shown. It can be seen that there is a clear relationship between the attributes, high values of `semi-major axis` correspond to low values of `ang. distance` and vice versa. But the relationship is not functional, and is not well described by rules. Nevertheless, the minimum variance itemset has values close to zero in the areas where there is a lot of data points. Minimum variance patterns are thus capable of discovering, and describing groupings of data points which are otherwise hard to define.

The `sonar` *dataset.* We now turn our attention to the well known `sonar` dataset. Since our method is somewhat sensitive to outliers, we removed every record which contained a value more than 3 standard deviations from the mean for some attribute. An interesting pattern has been found between attributes 15 and 44, see Figure 4. We can see that high values of both attributes never occur together. The actual relationship is reminiscent of the second artificial dataset presented above. The correlation coefficient is only -0.114; based on it, the pattern would have most probably been missed by traditional correlation analysis. This situation is similar to 'holes in data' analyzed in [4] which are well approximated in our framework.

5 Performance Analysis

We now present performance evaluation of the minimum variance itemset mining algorithm. The default parameters were $d = 2$ and maximum of $r = 3$ attributes per itemset. We found this combination to be flexible enough to discover complex patterns, which are still reasonably easy to interpret.

We used three datasets for testing: the extrasolar planet and sonar datasets described above, and a large Physics dataset from the KDD Cup 2004 competition with 80 attributes and 50000 records.

The algorithm has been implemented in C. Figure 5 (left) shows the influence of the parameter d on computation time for various minimum variance

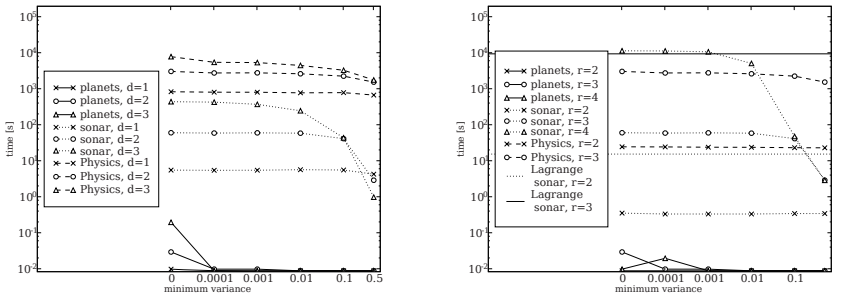

Fig. 5. Computation time vs. degree d (left) and max. itemset size r (right) for various minimum variance thresholds. Horizontal lines: Lagrange algorithm on sonar data.

thresholds. The parameter r is kept equal to the default value of 3. Figure 5 (right) shows the influence of the r parameter (d is kept equal to 2). Note that charts for $d = 2$ (left) and for $r = 3$ (right) in Figure 5 are identical since they correspond to the same parameter values. While performance of the algorithm is worse than for association rules in case of binary attributes (this is to be expected due to a much reacher structure of the data), the algorithm is practically applicable even for large datasets. It is interesting to see that, below a certain threshold, the minimum variance parameter has little influence on computation time.

We have also compared our approach with an equation discoverer Lagrange [9] (horizontal lines in Figure 5 (right)). The parameters were set such that it would discover polynomials of degree at most 2 involving at most 2 or 3 variables. Our approach was more than an order of magnitude faster than Lagrange. This is not surprising, as for every set of attributes Lagrange conducts an exhaustive search compared to a single relatively efficient eigenvalue computation in our case.

6 Conclusions and Future Research

A method for discovering arbitrarily complex relationships among numerical attributes has been presented. Its application yields itemsets and rules in the spirit of associations discovery. It has been shown experimentally that the approach does indeed produce interesting patterns, which capture various types of complex relationships present among the attributes. It is capable of finding patterns which would have been missed by standard polynomial regression analysis.

Future work is planned on increasing performance, *e.g.* by using bounds for eigenvalues to prune itemsets early.

References

1. Achtert, E., Böhm, C., Kriegel, H.-P., Kröger, P., Zimek, A.: Deriving quantitative models for correlation clusters. In: KDD, Philadelphia, PA, pp. 4–13 (2006)
2. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD, pp. 207–216 (1993)

3. Aumann, Y., Lindell, Y.: A statistical theory for quantitative association rules. In: Proc. of ACM-SIGKDD 1999, San Diego, CA, pp. 261–270 (1999)
4. Ku, L.-P., Liu, B., Hsu, W.: Discovering interesting holes in data. In: International Joint Conference on Artificial Intelligence (IJCAI 1997), pp. 930–935 (1997)
5. Besson, J., Robardet, C., De Raedt, L., Boulicaut, J.-F.: Mining bi-sets in numerical data. In: Džeroski, S., Struyf, J. (eds.) KDID 2006. LNCS, vol. 4747, pp. 9–19. Springer, Heidelberg (2007)
6. Brin, S., Motwani, R., Silverstein, C.: Beyond market baskets: Generalizing association rules to correlations. In: SIGMOD, pp. 265–276 (1997)
7. Calders, T., Goethals, B., Jaroszewicz, S.: Mining rank-correlated sets of numerical attributes. In: KDD, pp. 96–105 (2006)
8. Calders, T., Jaroszewicz, S.: Efficient auc optimization for classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 42–53. Springer, Heidelberg (2007)
9. Dzeroski, S., Todorovski, L.: Discovering dynamics: From inductive logic programming to machine discovery. J. of Intelligent Inform. Systems 4, 89–108 (1995)
10. Anderson, E., et al.: LAPACK Users' Guide. SIAM, Philadelphia (1999)
11. Fletcher, R.: Practical Methods of Optimization. Wiley, Chichester (2000)
12. Georgii, E., Richter, L., Rückert, U., Kramer, S.: Analyzing microarray data using quantitative association rules. Bioinformatics 21(2), ii1–ii8 (2005)
13. Healy, M.: Matrices for Statistics. Oxford University Press, Oxford (2000)
14. Jaroszewicz, S.: Polynomial association rules with applications to logistic regression. In: KDD, pp. 586–591 (2006)
15. Jaroszewicz, S., Korzeń, M.: Approximating representations for continuous data. In: SIAM'DM, pp. 521–526 (2007)
16. Karel, F.: Quantitative and ordinal association rules mining (qar mining). In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4251, pp. 195–202. Springer, Heidelberg (2006)
17. Wikipedia: Kepler's laws of planetary motion (retrieved February 26, 2007), `http://en.wikipedia.org/wiki/Kepler's_laws_of_planetary_motion`
18. Langley, P., Simon, H., Bradshaw, G., Zytkow, J.: Scientific Discovery. Computational Exploration of the Creative Process. MIT Press, Cambridge (1987)
19. Rückert, U., Kramer, S.: A statistical approach to rule learning. In: International Conference on Machine Learning (ICML 2006), Pittsburgh, PA, June 2006, pp. 785–792 (2006)
20. Rückert, U., Richter, L., Kramer, S.: Quantitative association rules based on half-spaces: An optimization approach. In: ICDM, pp. 507–510 (2004)
21. Schneider, J.: The extrasolar planets encyclopaedia, `http://exoplanet.eu`
22. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: ACM SIGMOD Conf. on Management of Data, pp. 1–12 (1996)
23. Steinbach, M., Tan, P.-N., Xiong, H., Kumar, V.: Generalizing the notion of support. In: KDD 2004, Seattle, WA, August 2004, pp. 689–694 (2004)
24. van Loan, C.F., Golub, G.H.: Matrix Computations. Johns Hopkins University Press (1996)
25. Zhang, H., Padmanabhan, B., Tuzhilin, A.: On the discovery of significant statistical quantitative rules. In: KDD, pp. 374–383 (2004)