Mining Atypical Groups for a Target Quantitative Attribute

Sylvie Guillaume
LIMOS Research Laboratory of Blaise Pascal University
Complexe scientifique des Cézeaux
63177 Aubière Cedex - France
sylvie.guillaume@isima.fr

Florian Guillochon
LIMOS Research Laboratory of Blaise Pascal University
Complexe scientifique des Cézeaux
63177 Aubière Cedex - France
flo.guillochon@orange.fr

***Abstract*— An important task in data analysis is the understanding of unexpected or atypical behaviors in a group of individuals. Which categories of individuals earn the higher salaries or, on the contrary, which ones earn the lower salaries? We present the problem of how data concerning atypical groups can be mined compared with a target quantitative attribute, like for instance the attribute "salary", and in particular for the high and low values of a user-defined interval. Our search therefore focuses on conjunctions of attributes whose distribution differs significantly from the learning set for the interval's high and low values of the target attribute. Such atypical groups can be found by adapting an existing measure, the intensity of inclination. This measure frees us from the transformation step of quantitative attributes, that is to say the step of discretization followed by a complete disjunctive coding. Thus, we propose an algorithm for mining such groups using pruning rules in order to reduce the complexity of the problem. This algorithm has been developed and integrated into the WEKA software for knowledge extraction. Finally we give an example of data extraction from the American census database IPUMS.**

***Keywords*— Quantitative associations, interestingness measures, negative and positive associations.**

I. Introduction

An important task in data analysis is the understanding of unexpected or atypical behaviors in a group of individuals. Which categories of individuals earn the higher salaries or, on the contrary, which ones earn the lower salaries?

Our purpose is to detect automatically each group of individuals that stand out significantly from the behavior of the learning set for a quantitative attribute and in particular for the interval's high and low values given by the user. Our search therefore focuses on conjunctions of attributes whose distribution differs significantly from the learning set for the interval's high and low values of the target attribute.

A closely related area to our work is association rule mining [1]. Association rules are relations between attributes of the form *X*→*Y*. In market basket data, *X* and *Y* are items such as cider or pancakes and in categorical data, *X* and *Y* are attribute-value pairs such as income>20K€ or occupation=″manager″. The problem of discovering association rules can be broken down into two subproblems : (1) finding all sets of items (*itemsets*) or attribute-value pairs that have transaction support above a given minimum. These sets are called frequent itemsets. (2) generating for each frequent itemset, all rules that have minimum confidence. To detect these atypical groups, we look not for only frequent itemsets (*in order to extract interestingness groups*) but also for itemsets whose support is statistically surprising that is to say significantly low or, on the contrary, significantly high compared to what could be expected. In so doing, we add a new semantics to our associations : we seek itemsets whose support is significantly different from that of the learning set (*i.e. groups of individuals that stand out significantly from the behaviour of the learning set*). We propose a new objective measure for extracting surprising itemsets based on an existing measure : the intensity of inclination [2] used for mining ordinal association rules. This measure prunes out the transformation step of quantitative attributes (*i.e. the discretization step followed by the step of complete disjunctive coding*). Most of the extensions of association rules to quantitative data require a discretization of the quantitative attributes [3] [4] [5]. Srikant and Agrawal in [3] have proposed a technique for automatic discretization and for merging narrow intervals into wider ones. Zhang et. al. in [4] use clustering methods to improve the partitioning of quantitative attributes. However, these methods can only approximate the best rules and cannot decide with certainty which rules are true associations and which are just artefacts of discretization. Ludl and Widmer in [6] and Bay in [7] have proved that a discretization without taking the context into account i.e. their association with the other attributes can lead us to non-optimal solutions. Thus, Mehta and Parthasarathy in [8] have proposed a contextual discretization.

We present a novel approach that works directly on quantitative attributes, without the need for any discretization. A weight will be attributed for each transaction, and the closer the transaction corresponds to the given criterion, the higher this weight will be, as in the case for example of, high wage earners. Another approach that has similarities with our method is the technique of fuzzy sets because the attribution of a weight to each transaction can be compared with a degree of membership. Kuok et. al. in [9], Zhang in [10] and Subramanyam and Goswami in [11] uses the technique of fuzzy sets for mining quantitative rules. This technique also requires partitioning the set of values of quantitative attributes into intervals. However, a transaction can belong to more than

978-1-4244-1674-5/08 /$25.00 ©2008 IEEE

one interval thanks to a membership function valued in the real unit interval [0,1]. Our method doesn't require discretization and uses the quantitative attribute as a whole. It allows us to obtain a global and synthetic view of the behavior of attributes without going into greater detail. It is a fact that not all the mined associations are of equal interest for the user and we think that a general search is sufficient at first. This can then be followed by a closer examination of the association if the user so decides [12].

The proposed measure for discovering atypical groups is a statistical measure, which makes a comparaison with the expected situation and then evaluates whether this difference is significant. Brin et. al. in [13] have introduced a variation of categorical association rules based on relating to associations as statistically interesting correlations. They have used the chi-square test of independence. The difference between using the intensity of inclination test and a chi-square test of independence is that the chi-square test evaluates the mean of differences between observed and expected situations for each value of the attribute, whereas the intensity of inclination test focuses on a given zone of the attribute as for example the high values.

The main contributions of our paper are as follows :

1. We propose an algorithm to mine a new semantics of itemsets : conjunctions of attributes verifying a significantly high support or, on the contrary, significantly low for a zone-defined by the user discarding the discretization and complete disjunctive coding steps for quantitative variables.

2. We propose a new interestingness measure for sets of qualitative and quantitative attributes which is an adaptation form of the intensity of inclination.

3. We present an evaluation of our proposed solutions on IPUMS to show their feasibility. Federal Census data is a difficult data set for most mining algorithms because there are many frequent and long itemsets.

The remainder of the paper is organized as follows. In *section 2* we define precisely the meaning of these atypical groups and in *section 3*, we present the interestingness measure and its adaptation which allows us to extract these groups. In *section 4*, we introduce two criteria which allow us to extract interestingness atypical groups, criteria which are then used for decreasing the complexity of the problem as described in *section 5* where we expose the algorithm. In *section 6*, our technique is evaluated using the American census database IPUMS and we conclude with a summary and further work.

II. Atypical Groups

In this section, we define the meaning of atypical groups, that is to say groups which are significantly over-represented or, on the contrary, significantly under-represented for a zone of the user-defined target attribute.

It seemed interesting to us to give the user the possibility of studying this quantitative attribute over a particular interval for several reasons. First, it permits us to remove the undefined values associated with this attribute, due to typing errors or to indicate that the value can not be filled in. Then, it permits us to remove the exceptional people that present very high or very low values for the target variable and whose presence might bias the results. For instance, if doing a study concerning salaries, it is preferable to remove the extremely wealthy individuals, like for example the author of the tales of a famous young wizard. Finally, it enables us to study a particular segment of the learning set, as for instance a study about average wage earners.

Then, our study focuses on a quantitative target attribute Z and more particularly on the zone of interest $Z = [z_1, z_2]$.

Let M be a conjunction of attributes called itemsets. Two kinds of itemsets are of interest : (1) qualitative itemsets X, that is to say itemsets only composed of attributes where a complete disjunctive coding step has been successfully completed (*i.e. attribute-value pairs such as occupation = ″manager″*) and (2) quantitative itemsets XY, that is to say qualitative itemsets X associated with a conjunction of quantitative attributes Y, quantitative attributes where no transformation step has been carried out. We also assimilate quantitative attributes Y (*quantitative attributes where no transformation step has been carried out*) with quantitative itemsets. In this paper, we are interested in associations between itemsets M and our target attribute $Z = [z_1, z_2]$. We call targeted association, the conjunction between itemset M and our target interval $Z = [z_1, z_2]$. We note this targeted association as : ZM in order to simplify its writing. When the itemset M=X is qualitative, we will say that the targeted association ZX is qualitative, and when the itemset M=XY is quantitative, we will say that the targeted association ZXY is quantitative.

When we calculate the support of an itemset M [14], each transaction verifying the itemset has the same importance. We are interested in targeted associations ZM, where no transformation step has been carried out with the attribute Z. We would like to discover categories of subjects that have a significantly high support or, on the contrary, a significantly low support, for the two zones of the interval of Z : the high and low values of this interval. To reach this goal, a weight has been attributed to each transaction : the closer the transaction corresponds to the expected criterion, the greater the weight is. For example in the case of ″*earn a high income*″ the closer individuals are to verifying a high value for the attribute ″*wage*″, the greater the weight will be. Thus, we propose two new support measures : a positive support which is interesting for the high values of the interval of Z and a negative support which is interesting for the low values of the interval of Z.

Let Ω be the learning set and e_i be a transaction (*or individual*) of the learning set. Let $C = (Z = [z_1, z_2])_{e_i \in \Omega}$ be the set of transactions verifying a value for the attribute Z between z_1 and z_2 and $(X)_{e_i \in \Omega}$ be the set of transactions verifying the qualitative itemset X. Let z_i be the value taken by the transaction e_i for the target attribute Z and x_i be the value taken by the transaction e_i for the qualitative attribute X ($x_i = 1$ *if the transaction* e_i *verifies the itemset* X, $x_i = 0$ *otherwise*).

Definition 1. The **positive support** $\text{supAbs}(ZX, z_2)$ of targeted qualitative association ZX is the weighted absolute support of the itemset X in the set C where the closer the value for transactions in Z to z_2, the greater the weight is.

$$\text{supAbs}(ZX, z_2) = \sum_{e_i \in C \cap (X)_{e_i \in \Omega}} (z_i\text{-}z_1) = \sum_{e_i \in C} (z_i - z_1) x_i$$

Definition 2. The **negative support** $\text{supAbs}(ZX, z_1)$ of targeted qualitative association ZX is the weighted absolute support of the itemset X in the set C where the closer the value for transactions in Z to z_1, the greater the weight is.

$$\text{supAbs}(ZX, z_1) = \sum_{e_i \in C \cap (X)_{e_i \in \Omega}} (z_2\text{-}z_i) = \sum_{e_i \in C} (z_2 - z_i) x_i$$

In the case of a targeted quantitative association ZXY, we have additional information : values taken by transactions of the set $(X)_{e_i \in \Omega} \cap C$ for this quantitative attribute Y. We are also interested in the high and low values taken by this attribute Y because we are not seeking to perform a transformation step on quantitative attributes. These concepts of negative and positive support lead us respectively to two new notions : support (*positive or negative*) for the low values of Y and support (*positive or negative*) for the high values of Y. Unlike the target attribute Z where the user can specify an interest zone $[z_1, z_2]$, we cannot define a study area for the other quantitative attributes Y associated with Z.

Definition 3. The **positive support** of a targeted quantitative association ZXY is the weighted absolute support of itemset XY in the set C where greater weight is given not only to transactions which have a value for Z close to z_2 but also to those which have

- a high value for Y (*targeted association ZXY+*),
- a low value for Y (*targeted association ZXY-*).

$$\text{supAbs } (ZXY+, z_2) = \sum_{e_i \in C \cap (X)_{e_i \in \Omega}} (z_i - z_1)(y_i - y_{\min})$$

$$\text{supAbs } (ZXY-, z_2) = \sum_{e_i \in C \cap (X)_{e_i \in \Omega}} (z_i - z_1)(y_{\max} - y_i)$$

with y_i being the value taken by transaction e_i for attribute Y. In the case where Y is a conjunction $Y_1..Y_j..Y_h$ of quantitative attributes, y_i is equal to the sum of normalized values taken by each quantitative attribute Y_j (*see definition of intensity of inclination in section 3*).

Definition 4. The **negative support** of targeted quantitative association ZXY is the weighted absolute support of itemset XY in the set C where greater weight is given not only to transactions which have a value for Z close to z_1 but also to those which have

- a high value for Y (*targeted association ZXY+*),
- a low value for Y (*targeted association ZXY-*).

$$\text{supAbs } (ZXY+, z_1) = \sum_{e_i \in C \cap (X)_{e_i \in \Omega}} (z_2 - z_i)(y_i - y_{\min})$$

$$\text{supAbs } (ZXY-, z_1) = \sum_{e_i \in C \cap (X)_{e_i \in \Omega}} (z_2 - z_i)(y_{\max} - y_i)$$

Our goal is to find all G groups (*qualitative and quantitative*) where one absolute supports s_0 = supAbs(ZM,z) (*see definitions 1 to 4*) differs significantly from the expected support that is to say from the calculated support under an assumption of independence between itemset M and the interest zone of the target attribute Z. By atypical group we mean, any group which has an absolute support that differs significantly from the expected one.

Let S be the random variable whose s_0 is an absolute support and α be the type I error (*also known as an error of the first kind, which is the error of rejecting a null hypothesis when it is actually true*).

Definition 5. A group G (*qualitative or quantitative*) is **atypical** if one of its absolute supports s_0 differs significantly from the expected one, that is to say when one of these two conditions is verified :

condition 1 : $Pr(S \leq s_0) \geq 1 - \alpha$

condition 2 : $Pr(S \leq s_0) \leq \alpha$

According to the verified condition, we will speak of positive atypical group (*condition 1*) or negative atypical group (*condition 2*).

Now, we will define the measure for extracting these atypical groups. It is a new measure based on an existing measure : intensity of inclination.

III. Measure

In this section we first remind the reader of the definition of intensity of inclination, a measure allowing implications between conjunctions of quantitative variables to be mined [2] and then we show how this measure has been adapted and extended to detect atypical groups.

A. Intensity of Inclination

In this section we give a more general definition to the attribute Z : it is a conjunction of quantitative attributes. Let Z and Y be respectively two conjunctions of p and q quantitative attributes. We suppose that $Z = Z_1, .., Z_p$ and $Y = Y_1, .., Y_q$, where $Z_1, .., Z_p, Y_1, .., Y_q$ are quantitative attributes taking values z_{1_i}, .., z_{p_i}, y_{1_i}, .., y_{q_i} ($i \in \{1..N\}$) respectively in intervals $[z_{1_{\min}} .. z_{1_{\max}}]$, .., $[z_{p_{\min}} .. z_{p_{\max}}]$, $[y_{1_{\min}} .. y_{1_{\max}}]$, .., $[y_{q_{\min}} .. y_{q_{\max}}]$.

Intensity of inclination evaluates whether the number of transactions not strongly verifying the rule $Z \rightarrow Y$ (*i.e. the number of transactions verifying simultaneously a high value for each attribute Z_1, .., Z_p and a low value for each attribute Y_1, .., Y_q*) is significantly small compared to the expected number of transactions under the assumption that Z and Y are independent. These transactions that do not strongly verify the rule are called negative transactions.

Let z_i and y_i be respectively values taken by attributes Z and Y in the database Ω for transaction e_i ($e_i \in \Omega$) and let z_{min} and y_{max} be respectively the minimum and maximum values taken by variables Z and Y.

The number t_0 of negative transactions, or raw measure of non-inclination, is defined by:

$$t_o = \sum_{i=1}^{N} (z_i\text{-}z_{\min})(y_{\max}\text{-}y_i) \quad \text{with}$$

$$z_i = \sum_{j=1}^{p} z'_{j_i} \text{ , } z_{\min} = \sum_{j=1}^{p} z'_{j_{\min}} \text{ , } y_i = \sum_{k=1}^{q} y'_{k_i} \text{ , } y_{\max} = \sum_{k=1}^{q} y'_{k_{\max}} \text{ ,}$$

$$z^{'}_{j_i} = \frac{z_{j_i} - \mu_{Z_j}}{\sigma_{Z_j}} \ (j \in \{1..p\}), \quad y^{'}_{k_i} = \frac{y_{k_i} - \mu_{Y_k}}{\sigma_{Y_k}} \ (k \in \{1..q\})$$

Let μ_{Z_j} and μ_{Y_k} be respectively the means of attributes Z_j ($j \in \{1, .., ,p\}$) and Y_k ($k \in \{1, .., ,q\}$) and let σ_{Z_j} and σ_{Y_k} be respectively the standard deviations of Z_j and Y_k.

The random variable T, whose t_0 is an observed value, can be approximated asymptotically by a normal distribution N (μ, σ) with $\mu = N\,(\mu_Z - z_{min})(y_{max} - \mu_Y)$ and $\sigma^2 = N\,[v_Z\, v_Y + v_Y\,(\mu_Z - z_{min})^2 + v_Z\,(y_{max} - \mu_Y)^2]$.

The means and variances of attributes Z and Y are given by the following expressions :

$$\mu_Z = \sum_{j=1}^{p} \mu_{Z_j} \text{ , } \mu_Y = \sum_{k=1}^{q} \mu_{Y_k} \text{ , } v_Z = \sum_{j=1}^{p} v_{Z_j} + 2\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p} \operatorname{cov}(Z_j, Z_{j'}) \text{ and}$$

$$v_Y = \sum_{k=1}^{q} v_{Y_k} + 2\sum_{k=1}^{q-1}\sum_{k'=k+1}^{q} \operatorname{cov}(Y_k, Y_{k'}) \text{ with } \operatorname{cov}(Z_i, Z_{i'}) = \mu_{Z_i Z_{i'}} - \mu_{Z_i}\mu_{Z_{i'}} \text{ .}$$

If the probability $Pr(\, T \leq t_o)$ of having a number inferior or equal to t_o is high, we can say that t_o is not significantly small because this occurrence can happen fairly frequently and then this implication $Z \rightarrow Y$ is not relevant.

To evaluate this implication in increasing order, the measure $\varphi\,(Z \rightarrow Y) = 1 - F(t_o) = Pr(T > t_o)$ has been retained where F is the cumulative distribution of T. Then, the implication $Z \rightarrow Y$ can be admitted with a level of confidence ($1\text{-}\alpha$) if and only if $Pr(T \leq t_o) \leq \alpha$ or $Pr(T > t_o) \geq 1\text{-}\alpha$.

The intensity of inclination is given by :

$$\varphi(Z \rightarrow Y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{t_0}^{+\infty} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt$$

Thus the intensity of inclination evaluates the *"smallness"* of the number of negative transactions as compared with independence. To do this, the measure calculates the sum of weights for all transactions in the database, weight being given by the raw measure of non-inclination. This weight is greatest for transactions verifying $Z = z_{max}$ and $Y = y_{min}$ (*corresponding perfectly to the concept of negative transactions*) and decreases until it is equal to zero for transactions verifying $Z = z_{min}$ or $Y = y_{max}$.

B. *Mining Atypical Groups*

First, in this section we show how the intensity of inclination has been used in a straightforward way for extracting qualitative atypical groups. However, for mining quantitative atypical groups, it is necessary to carry out a further measure.

Mining Qualitative Atypical Groups.

The intensity of inclination evaluates the *"smallness"* of the value t_0 as compared to what could be expected under the assumption that Z and Y are independent. Thus, we can use this measure to directly detect if the support *supAbs(ZX,z₁)* is significantly low with the set C as learning set. The *"smallness"* of the support will be significant if $\varphi(X \rightarrow Z) \geq (1\text{-}\alpha)$.

Now to know if this absolute support *supAbs(ZX,z₁)* is significantly high, we must verify that the complement to *1* of the value of the intensity of inclination $\varphi(X \rightarrow Z)$ can be admitted with a level of confidence ($1\text{-}\alpha$). A high support *supAbs(ZX,z₁)* will be significant if $1\text{-}\varphi(X \rightarrow Z) \geq (1\text{-}\alpha)$.

Table 1 summarizes formulas of the intensity of inclination for detecting qualitative atypical groups.

TABLE I. ADAPTATION OF THE INTENSITY OF INCLINATION FOR EXTRACTING QUALITATIVE ATYPICAL GROUPS

	Negative atypical	**Positive atypical**
supAbs(ZX,z₁)	$\varphi(X \rightarrow Z) \geq (1\text{-}\alpha)$	$1\text{-}\varphi(X \rightarrow Z) \geq (1\text{-}\alpha)$
supAbs(ZX,z₂)	$\varphi(Z \rightarrow 1\text{-}X) \geq (1\text{-}\alpha)$	$1\text{-}\varphi(Z \rightarrow 1\text{-}X) \geq (1\text{-}\alpha)$

Mining Quantitative Atypical Groups.

The use of the intensity of inclination for mining quantitative atypical groups is not straightforward because we have an extra attribute : the qualitative attribute X. Thus, we have to extend the definition of the intensity of inclination in order to use it.

Extension of the intensity of inclination

Let Y and Z be two conjunctions of quantitative attributes that take their values respectively in intervals [y_{min}, y_{max}] and [z_{min}, z_{max}]. Let X be a qualitative attribute. Let x_i, y_i and z_i be respectively values taken by transaction e_i for attributes X, Y and Z. Let N be the number of transactions in the learning set Ω. Let μ_X, μ_Y and μ_Z be respectively the means of attributes X, Y and Z and let v_X, v_Y and v_Z be respectively the standard deviations of attributes X, Y and Z.

The raw measure is defined by:

$$t'_0 = \sum_{i=1}^{N} (z_i\text{-}z_{\min})(y_{\max}\text{-}y_i)\ x_i = \sum_{i=1}^{N} (z_i\text{-}z_{\min})\ u_i$$

The random variable T', whose t'_0 is an observed value, can be approximated asymptotically by a normal distribution N (μ', σ') with
$\mu' = N\,(\mu_Z - z_{min})\mu_U$ and $\sigma'^2 = N\,[v_Z\, v_U + v_U\,(\mu_Z - z_{min})^2 + v_Z\,(\mu_U)^2]$.
with $\mu_U = (y_{max} - \mu_Y)\mu_X$
and $v_U = [v_X\, v_Y + v_X\,(y_{max} - \mu_Y)^2 + v_Y\,(\mu_X)^2]$.

If the probability $Pr(\, T' \leq t'_o)$ of having a number inferior or equal to t'_o is high, we can say that t'_o is not significantly small because this occurrence can happen fairly frequently.

In order to evaluate the *"smallness"* of this difference in increasing order, the measure $\varphi'(Z \rightarrow XY) = Pr(T' > t'_o)$ can be used. Thus, the *"smallness"* of this difference can be admitted with a level of confidence ($1\text{-}\alpha$) if and only if $Pr(T' \leq t'_o) \leq \alpha$ or $Pr(T' > t'_o) \geq 1\text{-}\alpha$.

Thus, the extended intensity of inclination is given by :

$$\varphi'(Z \rightarrow XY) = \frac{1}{\sigma\sqrt{2\pi}} \int_{t'_0}^{+\infty} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt$$

and the *"smallness"* of this difference is significant if $\varphi'(Z \rightarrow XY) \geq 1\text{-}\alpha$.

Application of the extension of the intensity of inclination
We can use this extension of the measure to directly detect if the support *supAbs*(ZXY-,z_2) is significantly low with the set C as learning set.

$$\text{supAbs }(ZXY-,z_2)=\sum_{e_i\in C}(z_i-z_1)(y_{\max}-y_i)x_i$$

The *″smallness″* of this support is significant if
$\varphi'(Z\rightarrow XY)\geq(1-\alpha)$.
In the same way as for qualitative atypical groups, this support will be significantly high if the complement to *1* of the value of the intensity of inclination $\varphi'(Z\rightarrow XY)$ can be admitted with a level of confidence ($1-\alpha$). A high support *supAbs*(ZXY-,z_2) will be significant if $1-\varphi'(X\rightarrow Z)\geq(1-\alpha)$.

Table 2 summarizes formulas of the extension of the intensity of inclination for detecting quantitative atypical groups.

TABLE II. FORMULAS OF THE EXTENSION OF THE INTENSITY OF INCLINATION FOR DETECTING QUANTITATIVE ATYPICAL GROUPS

	Negative atypical	**Positive atypical**
supAbs(ZXY-,z_2)	$\varphi'(Z\rightarrow XY)$	$1-\varphi'(Z\rightarrow XY)$
supAbs(ZXY+,z_2)	$\varphi'(Z\rightarrow X(y_{max}+y_{min}-Y))$	$1-\varphi'(Z\rightarrow X(y_{max}+y_{min}-Y))$
supAbs(ZXY-,z_1)	$\varphi'(X(y_{max}+y_{min}-Y)\rightarrow Z)$	$1-\varphi'(X(y_{max}+y_{min}-Y)\rightarrow Z)$
supAbs(ZXY+,z_1)	$\varphi'(XY\rightarrow Z)$	$1-\varphi'(XY\rightarrow Z)$

IV. INTERESTINGNESS ATYPICAL GROUPS

We have exposed the concept of atypical groups and how we have adapted and extended the intensity of inclination to extract them. Now we define two criteria that allow us to discover interestingness atypical groups.

The first criterion allows us to extract noteworthy atypical groups (*frequent atypical groups*) and the second one ensures that each extracted atypical group brings new information (*informative atypical groups*).

Frequent atypical groups.

The first criterion, *criterion 1*, ensures that the atypical group is noteworthy that is to say composed of a minimum number of transactions. By analogy with the definition of frequent itemsets (Agrawal et al., 1996), we call these groups, frequent atypical groups.

Criterion 1. Let s_1 be the user-defined threshold called the minimum support. An atypical group G is said to be frequent when, the associated qualitative itemset X is frequent, that is to say if Pr(X/Z) ≥ s_1.

Informative atypical groups.

The second criterion, *criterion 2*, ensures that each extracted atypical group is informative that is to say it gives new information that may not occur with a more general atypical group.

In order to define this criterion, we need to introduce the concept of general or super-group.

Definition 6. Let M and M' be respectively the associated itemsets with groups G and G'. Group G is a super-group of group G' if $M\subset M'$ that is to say if the itemset M is included in the itemset M'. We also can say that group G' is a subgroup of G.

Thus, for instance, the itemset M=(*sex*=*″female″*) is included in the itemset M'=(*sex*=*″female″*)∧(*occupation*=*″manager″*).

From now on, each name of a group followed by an apostrophe means that it is a subgroup, i.e. a group whose associated itemset is composed of at least two attributes.

Extracting uninteresting positive atypical groups G' (*respectively negative*) can occur when one of its super-groups is highly over-represented (*respectively highly under-represented*) for the studied zone of the target attribute. We give the example of professional people, who have studied for many years, and therefore are highly over-represented in the zone of high values of the attribute *″education″*. Any itemset associated with the itemset occupation=*″professional″*, as for instance the itemset sex=*″female″*, will add no new information because this new group G' (*professional women*) is a subset of professional people. We can formalize the discovery of informative atypical groups as follows :

Definition 7. Let s_2 be a user-defined maximum support.

The positive atypical group (*respectively negative*) G' is **informative** as compared with a zone z of the target attribute Z if none of its positive G super-groups (*respectively negative*) are highly over-represented (*respectively highly under-represented*) in this same zone, that is to say don't have a support which is superior (*respectively inferior*) to a given threshold s_2 :

Positive group : $\forall\ M\subset M'$ supAbs(ZM, z) $\leq s_2$
Negative group : $\forall\ M\subset M'$ supAbs(ZM, z) $\geq s_2$

According to the studied zone, we have two possible supports : supAbs(ZM,z_1) for the zone of low values of the target interval and supAbs(ZM,z_2) for the zone of high values. When itemset M is a quantitative itemset (M=XY), we have two new supports : the support supAbs(ZXY+, z) for high values for Y and the support supAbs(ZXY-, z) for low values for Y.

After defining these two criteria which allow us to extract interestingness groups, we expose how we have used them for pruning the search area for itemsets.

V. ALGORITHM

In this section, we present our algorithm for mining positive interestingness atypical groups for the interval's high values of the target attribute. We can easily transpose this search to the following: (1) mining positive interestingness atypical groups for the interval's low values and (2) mining negative interestingness atypical groups for the interval's high and low values.

Figure 1 shows our algorithm for mining positive interestingness atypical groups. The algorithm is based on the *Apriori* algorithm [14] and more particularly on the first step : extraction of frequent itemsets. The input for our algorithm is : the database where a complete disjunctive coding has been realized on the qualitative attributes, the target quantitative attribute Z, the interval $[z_1, z_2]$, the type I error α, the minimum and maximum support. The algorithm returns the set of positive interestingness atypical groups for the user-defined zone.

We use an Apriori-like algorithm [14] and our algorithm visits the lattice of itemsets in a level-wise fashion and uses the anti-monotone property of support and *property 1* described below to eliminate large portions of the search area. Like the Apriori algorithm, the complexity of our algorithm has linear complexity with the same number of multiple passes on the database but exponential complexity with respect to the number of attributes.

This *property 1* uses *criterion 2* presented in *section 4*, which tells us that if itemset M is highly frequent (*respectively highly unfrequent*) for a given zone z then each group G' associated with an over-itemset M' will be not informative.

```
Algorithm
Input : data, target quantitative attribute $Z$, interval $[z_1, z_2]$, type I error $\alpha$, minimum support $s_1$ and maximum support $s_2$.
Output : set of positive interestingness atypical groups or set of itemsets associated with groups.
BEGIN
//(1) Calculation of frequent qualitative itemsets $X_i$ of level $1$
$LC_1=\{X_i$ / support$(X_i) \geq s_1$ and $|X_i|=1\}$
//(2) Calculation of itemsets $M$ whose associated $G(M)$ group is a positive atypical group
$LAC_1=\{X_i$ / $X_i \in LC_1$ and $1-\varphi(Z \rightarrow 1-X_{i,}) \geq (1-\alpha)\}$
$LAQ_1=\{Y_i(+/-)$ / $1-\varphi(Z \rightarrow Y_i) \geq (1-\alpha)$ or
                         $1-\varphi(Z \rightarrow (y_{max}+y_{min}-Y_i)) \geq (1-\alpha)\}$
//(3) Detection of highly frequent itemsets $X_i$ and $Y_i(+/-)$
$LPC_1=\{X_i$ / $X_i \in LAC_1$ and supAbs$(ZX_i, z_2) \geq s_2\}$
$LPQ_1=\{Y_i(+/-)$ / $Y_i(+/-) \in LAQ_1$ and
                         supAbs$(ZY_i(+/-), z_2) \geq s_2\}$
$k=2$
While $LC_{k-1} <> \emptyset$ do
  //(4) Generation of candidate level $k$ itemsets
  $CC_k=\{X_i$ / $|X_i|=k\}$ from $LC_{k-1} \setminus LPC_{k-1}$
  $LQ_k=\{X_iY_i(+/-)$ / $|X_i| \leq k-1$, $1 \leq |Y_i(+/-)| \leq k$ and $|X_iY_i(+/-)| = k\}$
           from $LC_{m-1} \setminus LPC_{m-1}$ $(m \leq k)$ and $LAC_n \setminus LPQ_n$ $(n \leq k)$
  //(5) Calculation of frequent qualitative level $k$ itemsets
  $LC_k=\{X_i$ / $X_i \in CC_k$ and support$(X_i) \geq s_1\}$
  //(6) Calculation of itemsets $M$ whose associated $G(M)$ group is a positive atypical group
  $LAC_k=\{X_i$ / $X_i \in LC_k$ and $1-\varphi(Z \rightarrow 1-X_{i,}) \geq (1-\alpha)\}$
  $LAQ_k=\{X_iY_i(+/-)$ / $X_iY_i(+/-) \in LQ_k$ and $1-\varphi'(Z \rightarrow X_iY_i) \geq (1-\alpha)$
                         or $1-\varphi'(Z \rightarrow X_i(y_{max}+y_{min}-Y_i)) \geq (1-\alpha)\}$
  //(7) Detection of highly frequent itemsets
  $LPC_k=\{X_i$ / $X_i \in LAC_k$ and supAbs$(ZX_i,+) \geq s_2\}$
  $k=k+1$
End While
Return $\bigcup_{i=1..k} (LAC_i \cup LAQ_i)$
END
```

Figure 1. Algorithm for mining positive interestingness atypical groups for the interval's high values of the target quantitative attribute Z.

Property 1.

Positive group : **if** supAbs(ZM, z) $\geq s_2$ **then** $\forall$ M' $M \subset M'$ $G'(M')$ is not informative

Negative group : **if** supAbs(ZM, z) $\leq s_2$ **then** $\forall$ M' $M \subset M'$ $G'(M')$ is not informative

First, the algorithm searches for more general positive atypical groups that is to say groups whose associated itemset is composed of one attribute (*see steps 1 to 3*). *Step 1* extracts frequent qualitative itemsets. *Step 2* mines positive atypical groups from frequent qualitative and quantitative attributes. *Step 3* detects highly frequent itemsets (*see criterion 2 in section 4*) which will then be discarded from the set of frequent itemsets used for generating candidate itemsets.

Then, *steps 4* to *7* extract interestingness positive atypical subgroups and these four steps will be repeated for each *k level* (*a k level represents the searching for itemsets composed of k attributes*) until the set of lower level frequent qualitative itemsets is equal to the empty set (*because we will not be able to generate candidat itemsets*). *Step 4* generates simustaneously : (1) candidate qualitative itemsets (*as in the Apriori algorithm*) from lower level frequent itemsets deprived of highly frequent itemsets, and (2) candidate quantitative itemsets composed of at least one quantitative attribute not highly frequent and one frequent qualitative itemset also not highly frequent. *Step 5* calculates frequent qualitative itemsets from candidate itemsets found in *step 4*. *Step 6* extracts interestingness positive atypical groups from frequent qualitative itemsets found in *step 5* and from candidate quantitative itemsets found in *step 4*. *Step 7* detects highly frequent itemsets in the set of qualitative itemsets in order to discard them from the set of frequent qualitative itemsets which generate upper level candidate itemsets.

Our algorithm has been implemented in Java and integrated into the WEKA (***W**aikato **E**nvironnement for **K**nowledge **A**nalysis*) software for knowledge extraction [15].

VI. EXPERIMENTATIONS

Our technique is evaluated using the American census database IPUMS 99 (*Integrated Public Use Microdata Series*). Federal Census data is a difficult data set for most mining algorithms because there are many frequent and long itemsets. The IPUMS data is available from the UCI KDD archive [16].

The IPUMS data consists of *88 443* transactions described by *61* attributes including *29* quantitative attributes. The complete disjunctive coding step of qualitative attributes has returned *772* attributes. We have focused our study on the target quantitative attribute : *"wage"*. Values taken by this attribute are in the interval [*0, 999 999*]. However *23,98*% of transactions verify the value *999 999*. This percentage doesn't reflect reality : it is a value by default for transactions where this attribute *"wage"* has no significance. That is why, we have choosen the following interval [*0*, *195 516*] to identify atypical groups : *195 516* is the value immediately inferior to *999 999*. The parameters were set as follows : type I error was equal to *0.05* and the minimum support was equal to *0.01*.

These associations reveal that atypical groups in the class of persons earning a high wage are as follows :

Qualitative atypical groups :

- Men are over-represented, on the contrary, women are under-represented.

- However, men who belong to the following three occupational categories : ″*truck and tractor drivers*″ and ″*operative and kindred workers*″ are under-represented, whereas women who belong to the following three occupational categories : ″*teachers*″, ″*managers, officials and proprietors*″ and ″*stenographers, typists and secretaries*″ are over-represented.

- We also learn that men who are divorced and born in either ″*California*″ or ″*Central America*″ are under-represented whereas women who are divorced and born in ″*California*″ are over-represented.

- ″*never married*″ men, born in ″*Mexico*″ and living in ″*California*″ are under-represented.

- Individuals whose educational attainment is ″*1 to 3 years of college*″ or ″*4+ years of college*″ are over-represented whereas the other levels (″*none or preschool*″, ″*grade 1-4*″, ″*grade 5-8*″, ″*grade 9*″, ″*grade 10*″ and ″*grade 11*″) are under-represented.

- However, we learn that individuals whose educational attainment is ″*1 to 3 years of college*″ and either ″*widowed*″ or ″*single*″ or born in ″*Mexico*″ are under-represented whereas individuals whose the educational attainment is ″*grade 11*″ and either ″*veteran status= yes*″ or ″*labor force status = yes*″ or ″*employement status = employed*″ are , on the contrary, over-represented.

Quantitative atypical groups :

- The higher the ″*total personal income*″, the higher the wage.

- The higher the ″*number of own siblings in household*″, the lower the wage.

- Households with a ″*number of children under age 5*″ near to 5 are over-represented.

- The older the head of the family, the higher the wage when he was born in ″*Missouri*″, whereas the older the head of the family, the lower the wage when he was born in ″*Central America*″.

- If the household is a home-owner, and the head of the family was born in ″*Mexico*″, and if the ″*total personal income*″ of the household is either low, or high, in both cases, the two corresponding groups (*low and high total income*) are under-represented.

VII. CONCLUSION AND FURTHER WORK

In this paper, we have proposed a technique for extracting interestingness atypical groups for a target quantitative attribute, and in particular for the high and low values of a user-defined interval. This method reveals a new semantics of itemsets : conjunctions of attributes verifying a significantly high support or, on the contrary, significantly low for a user-defined zone that is to say, conjunctions of attributes having a behavior which is different from the learning set. The proposed measure for mining these associations discards the discretization and complete disjunctive coding steps for quantitative variables and has the following advantages : (1) it eliminates errors associated with a priori discretization and (2) it provides a global view of associations and thus the generated knowledge for this kind of attribute is not parsed.

The user defines a study interval for the target quantitative attribute and it could be interesting to propose a technique which automatically detects the best zones where the group is particularly atypical. Also, another input parameter could be proposed : a reference set, not necessary the learning set, and all groups having a similar or different behavior could be mined.

ACKNOWLEDGMENT

We thank CAROL FYNN for her invaluable assistance with the writing of this paper in English.

REFERENCES

[1] Agrawal R., Imielinski T. and Swami A., Mining associations between sets of items in massive database, In Proceedings of the *ACM SIGMOD International Conference on Management of Data*, 207-216, 1993.

[2] Guillaume S., Discovery of Ordinal Association Rules, *6th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD'02), 322-327, Taipei, Taiwan, 2002.

[3] Srikant, R., Agrawal, R. (1996) Mining Quantitative Association Rules in Large Relational Tables, *ACM-SIGMOD International Conference Management of Data*, Montréal, Canada, 1996.

[4] Zhang Z., Lu Y. and Zhang B., An effective partitioning-combining algorithm for discovering quantitative association rules, *Proc. of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining,* 1997.

[5] Wijsen J. and Meersman R., On the complexity of mining quantitative association rules, *Data Mining and Knowledge Discovery*, 2(3):263-281, 1998.

[6] Ludl M.C. and Widmer G., Relative Unsupervised Discretization for Association Rule Mining, Proc. 4th European Conference Principles and Practice of Knowledge Discovery in Databases, 148-158, 2000.

[7] Bay S.D., *Multivariate Discretization for Set Mining*, Knowledge and Information Systems, vol.3, n°4, 491-512, 2001.

[8] Mehta S. and Parthasarathy S., *Toward Unsupervised Correlation Preserving Discretization*, IEEE Transactions on Knowledge and Data Engineering, vol.17, n°9, 1174-1185, 2005.

[9] Kuok, C.M. Fu, A., and Wong, M.H., Mining Fuzzy Association Rules in Databases, *ACM SIGMOD Record*, 41-46, 1998.

[10] Zhang W., Mining Fuzzy Quantitative Association Rules, *11th IEEE International Conference on Tools with Artificial Intelligence,* 1999.

[11] Subramanyam R.B.V. and Goswami A. (2006), Mining Fuzzy Quantitative Association Rules, *Expert Systems*, Vol. 23, N°4, 212-225.

[12] Guillaume S., Ordinal Association Rules towards Association Rules, In *proceedings of the 5th International Conference on Data Warehousing and Knowledge Discovery* (DaWaK 2003), 3-5 September 2003, p. 161-171, Prague, Czech Republic, ISBN 3-540-40807-X.

[13] Brin, S. Rastogi, R. and Shim, K., Mining Optimized Gain Rules for Numeric Attributes, *IEEE transactions on Knowledge and Data Engineering*, 324-338, 2005.

[14] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen & A.I. Verkamo – "Fast Discovery of Association Rules" - In Fayyad U.M., Piatetsky-Shapiro G., Smyth P. and Uthurusamy R. eds., Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press. Chapter 12, pp. 307-328, 1996.

[15] Witten I.H. and Frank E., *Data Mining, practical machine learning tools and techniques with Java implementations*, Morgan Kauffman, ISBN 0-12-088407-0, 2005.

[16] Bay S.D., The UCI KDD archive, [http://kdd.ics.uci.edu/], Irvine, CA : University of California, Department of Information and Computer Science, 1999.