Tight Optimistic Estimates for Fast Subgroup Discovery

Henrik Grosskreutz, Stefan Rüping, and Stefan Wrobel

Fraunhofer IAIS, Schloss Birlinghoven, 53754 St. Augustin, Germany
{henrik.grosskreutz,stefan.rueping,stefan.wrobel}@iais.fraunhofer.de

Abstract. Subgroup discovery is the task of finding subgroups of a population which exhibit both distributional unusualness and high generality. Due to the non monotonicity of the corresponding evaluation functions, standard pruning techniques cannot be used for subgroup discovery, requiring the use of optimistic estimate techniques instead. So far, however, optimistic estimate pruning has only been considered for the extremely simple case of a binary target attribute and up to now no attempt was made to move beyond suboptimal heuristic optimistic estimates. In this paper, we show that optimistic estimate pruning can be developed into a sound and highly effective pruning approach for subgroup discovery. Based on a precise definition of optimality we show that previous estimates have been tight only in special cases. Thereafter, we present tight optimistic estimates for the most popular binary and multi-class quality functions, and present a family of increasingly efficient approximations to these optimal functions. As we show in empirical experiments, the use of our newly proposed optimistic estimates can lead to a speed up of an order of magnitude compared to previous approaches.

1 Introduction

Subgroup discovery [Klö96, Wro97] is the task of finding subgroups of a population with high generality and distributional unusualness. It is a general approach that has shown to be useful in a variety of application scenarios, like medical consultation systems [ABP06], spatial analysis [KM02], marketing campaign planning [LCGF04], and also in contrast set mining tasks [KLGK07].

Unfortunately, if applied to real-world problems, subgroup discovery quickly results in excessive computation, due to its exponential dependency on the number of attributes. Different approaches have proposed to cope with that problem: While sampling based approaches [SW00] relax the task by allowing a certain degree of departure from the optimal solution and a (controllable) error probability, other approaches make use of sophisticated data structures [AP06] or heuristics [Klö02, LKFT04].

Another approach, proposed by Wrobel in [Wro97], is to prune the search space using so-called *optimistic estimates*. An optimistic estimate is a function that, given a subgroup s, provides a bound for the quality of every subgroup s' that is a refinement of s. Surprisingly, the use of optimistic estimates for fast

subgroup discovery has not yet been developed into a mature technology: until recently optimistic estimates have only been considered for the extremely simple case of a binary target attribute, and even in this case no attempt was made to move beyond suboptimal optimistic estimates.

In this paper, we investigate the question whether the optimistic estimates considered so far provide bounds that are, in some sense, optimal. To this end we provide a formal definition of *tight* optimistic estimates, that is optimistic estimates that are as conservative as possible wrt. the information at hand, namely the size of a subgroup and its distribution over different classes. Using this definition, we show that the optimistic estimate proposed in [Wro97] is not tight. Thereafter, we present new tight optimistic estimates for some of the most common quality functions. We also present a family of increasingly efficient approximations to these optimal functions. While these optimistic estimates are not tight, they have the advantage that they are simpler to calculate.

Summarizing, the main contributions of this paper are thus (i) the formal definition of tightness, (ii) new, tight optimistic estimates for some of the most common quality functions, and (iii) a family of increasingly efficient approximations to these optimal functions. In an experimental section, we show that our results are not only interesting from a theoretical point of view, but also have a significant impact on the performance of subgroup discovery algorithms.

The paper is organized as follows: In section 2, we define the basics of the subgroup discovery task. In section 3, we provide our definition of tight optimistic estimates; thereafter, we present and prove new (tight) optimistic estimates. Section 4 contains the experiments, while section 5 concludes.

2   Preliminaries

In this section, we will introduce our terminology, formally define the problem of subgroup discovery, and motivate the concept of optimistic estimates.[1]

2.1   The Task of Subgroup Discovery

Let $DB = \{R_1, ..., R_N\}$ be a *database* or *dataset*, consisting of N rows, each built up from of $l+1$ values. We distinguish one attribute c, called the *class attribute* with domain $D(c) = \{c_1, ...c_m\}$, from the l ordinary attributes $\{a_1, a_2, ...a_l\}$ with domains $D(a_i) = \{v_{i,1}, ..., v_{i,m_i}\}$. Every database row R_j is an n-tuple $(v_{j,1}, ..., v_{j,l}, c_j)$, and we call c_j its class.

A *subgroup description sd* is a set of *terms* $\{t_1, ..., t_k\}$ where every term t_i is a constraint on an attribute, i.e. t_i has the form $(a_i = v_i)$, $v_i \in D(a_i)$. The *length* of the subgroup description is the number of terms it is built of. We call a subgroup description $sd' = \{t'_1, ..., t'_{k'}\}$ a *refinement* of a subgroup description $sd = \{t_1, ..., t_k\}$, denoted by $sd' \succ sd$, if $\{t_1, ..., t_k\}$ is a subset of $\{t'_1, ..., t'_{k'}\}$.

---

[1] In the following presentation, we assume that a dataset is provided as a single table. However, the concept of (tight) optimistic estimates also applies to the multi-relational setting involving joins over relations as considered in [Wro97].

Given a database DB and a subgroup description sd, the *subgroup extension* of sd on DB is the set of rows $R_j \in DB$ that satisfy all terms $t_i \in sd$. Please note that if sd' is a refinement of sd, i.e. $sd' \succ sd$, then for every database DB the subgroup extension for DB and sd' is a subset of the subgroup extension for DB and sd.

Given a set of rows $R = \{R_1, ...R_n\}$ (a database or subgroup extension), we call n its *size* and $\boldsymbol{p} = (p_1, ..., p_m)$, where p_i is the fraction of the rows of R of class i, its *class distribution*. Formally, $\boldsymbol{p}$ is defined as follows:

$$p_i := 1/n \times |\{r | r \in R \wedge class(r) = i\}|.$$

Figure 1 shows an example with hypothetical data, inspired from a medical domain. The rows represent medical prescriptions made by doctors. As class attribute, we consider the cost of the prescription. Beside this special attribute, the prescriptions contain the doctor's specialty and the information whether the doctor's practice is in an urban or a rural environment. In this example, $\{Specialty = Surgery\}$ and $\{Specialty =$

Cost	Specialty	Region
High	Surgery	Urban
Medium	Internal Med	Urban
Medium	Psychiatry	Urban
Medium	Internal Med	Rural
Low	Surgery	Rural
Low	Surgery	Rural

Fig. 1. Prescription example

$Surgery, Region = Urban\}$ are two subgroup descriptions, and the corresponding subgroups consist of the rows that fulfill these conditions. The size of the subgroup extension of $\{Specialty = Surgery\}$ is 3 and its probability distribution is $p_{high} = 0.33$, $p_{medium} = 0$, $p_{low} = 0.66$.

A *quality function* q is a mapping from $DB \times sd$ to the reals. Intuitively, a quality function expresses how "interesting" a subgroup is. Almost all quality functions considered in the literature only depend on some parameters of the subgroup and the database, in particular on the size n of the subgroup, the size N of the database, the class distribution $\boldsymbol{p}$ of the subgroup, and the class distribution $\boldsymbol{p_0}$ of the database. Table 1 summarizes some of the most prominent quality functions [Klö96, SW00]: the Piatetsky-Shapiro quality function dealing with the two-class case, and the Split, Gini and Person χ^2 quality functions for n-ary class attributes.[2]

The problem of *subgroup discovery* is defined as follows: Given a database DB, a quality function q, and a number k, determine the k subgroup descriptions with maximum quality. Or, put more formally: return a set of k subgroup descriptions G such that

$$\forall \text{ subgroup descriptions } sd : sd \notin G \Rightarrow q(DB, sd) \le q*,$$

where $q* = \min_{sd \in G} q(DB, sd)$.

---

[2] The notation and definitions used in other papers like [KLJ03, Wro97] sometimes slightly differ from ours, but are (factor-) equivalent. For example, the Gini-Quality is often expressed using the *generality* (i.e. n/N) of the subgroup. Other authors use a notation like $p(class | s)$ to denote the probability of a class in a subgroup.

Table 1. Common quality functions for subgroups

Name	Type	Definition
Piatetsky-Shapiro	2	$n(p - p_0)$
Split	N	$n \sum_i (p_i - p_{0\,i})^2$
Gini	N	$\frac{n}{N-n} \sum_i (p_i - p_{0\,i})^2$
Pearson's χ^2	N	$n \sum_i \frac{(p_i - p_{0\,i})^2}{p_{0\,i}}$

2.2 Optimistic Estimates and Their Use in Subgroup Discovery

Before we present the definition of optimistic estimates, we would like to motivate this concept by taking a look at possible algorithmic approaches to subgroup discovery. Given that the space of candidate subgroup descriptions can be considered as a tree with subgroup descriptions of length 1 at the first level, subgroup descriptions of length 2 at the next level and so on, one obvious approach to subgroup discovery is to perform some kind of search.

Of course, in this approach the size of the search space is exponential in the number of attributes and hence it is desirable to use some kind of pruning strategy. Unfortunately, unlike related tasks like frequent item mining where state-of-the art algorithms like FpGrowth [HPYM04] exploit the property of *monotonicity*, in subgroup discovery this property does not hold: Even if the subgroup description $a_1 = x$ does not have a sufficient quality, it is still necessary to consider its refinements. In fact, even if neither $a_1 = x$ nor $a_2 = y$ are interesting subgroups, $(a_1 = x, a_2 = y)$ might very well be interesting.

However, if we have already found k subgroups and we knew that all refinements s' of a subgroup s had a quality that is worse than that of all k subgroups found so far, we could safely prune that branch. What is needed to do so is an *optimistic estimate* for the refinements s' of s [Wro97]:

Definition 1. *An optimistic estimate $oe(s)$ for a given quality function q is a function that satisfies the following: $\forall$ subgroups $s, s'. s' \succ s \implies oe(s) \geq q(s')$.*

3 Tight Optimistic Estimates

In this section, we will present our definition of *tight* optimistic estimates. Thereafter, we will present new optimistic estimates for all quality functions in Table 1.

3.1 A Definition of Tightness with Respect to Probability and Size

The quality functions from Table 1 are all defined in terms of a few characteristics of the subgroup and the dataset, namely

- the distributions over the classes in the subgroup, denoted by $\boldsymbol{p}$;
- the size of the subgroup, denoted by n;
- the distributions over the classes in the dataset, denoted by $\boldsymbol{p_0}$; and
- the size of the dataset, denoted by N.

We will call such quality functions "probability/size quality functions" or "p/n quality functions". Formally, a p/n quality function is a function $q(\boldsymbol{p}, n, \boldsymbol{p_0}, N)$ from $[0,1]^c \times N \times [0,1]^c \times N$ to the reals (here, c is the number of classes). We call $\boldsymbol{p}$ and n the *parameters* of the subgroup and $\boldsymbol{p_0}$ and N the parameters of the overall population. Similarly, we will call optimistic estimates that only make use of these parameters "p/n optimistic estimates". Formally, a p/n optimistic estimate is a function from $[0,1]^c \times N \times [0,1]^c \times N$ to the reals such that

$$\forall \text{ subgroups } s, s'. \, s' \succ s \implies oe(\boldsymbol{p}(s), n(s), \boldsymbol{p_0}, N) \geq q(\boldsymbol{p}(s'), n(s'), \boldsymbol{p_0}, N).$$

Here, $\boldsymbol{p}(s)$ and $n(s)$ denote the class distribution and the size for subgroup s. In general, there are infinitely many optimistic estimates. We are interested in optimistic estimates that are as *conservative* as possible in the following sense:

Definition 2. *Given a quality function q and two optimistic estimates oe_1 and oe_2, we call oe_1 is more* conservative *than oe_2 if $\forall N, \boldsymbol{p_0}, n, \boldsymbol{p}. \, oe_1(\boldsymbol{p}, n, \boldsymbol{p_0}, N) \leq oe_2(\boldsymbol{p}, n, \boldsymbol{p_0}, N)$.*

The more conservative an optimistic estimate, the larger part of the search space can potentially be pruned: if we have already found k subgroups with a minimum quality $minQ$, then we can prune the branch of subgroups below s if and only if $oe(s) < minQ$. We will now formally define the notion of *tight* optimistic estimates, i.e. optimistic estimates that are as conservative as possible:

Definition 3. *An optimistic estimate oe for a quality function q is* tight *if for any population DB and any subgroup description sd there is a subset s' of the extension of sd on DB such that the quality of s' is equal to the optimistic estimate for sd on DB. Formally: the optimistic estimate oe is* tight *iff*

$$\forall DB, sd \, . \, \exists n', \boldsymbol{p'}. \, [n' \leq n \wedge n'\boldsymbol{p'} \preceq n\boldsymbol{p} \wedge oe(\boldsymbol{p}, n, \boldsymbol{p_0}, N) = q(\boldsymbol{p'}, n', \boldsymbol{p_0}, N)].$$

Here, $\boldsymbol{p_0}$ and $\boldsymbol{p}$ denote the probability distribution in DB, respectively in the extension of sd on DB, while N and n denote the size of DB respectively of the subgroup extension (actually, $\boldsymbol{p}, n, \boldsymbol{p_0}$ and N are functions of DB and s). $n'\boldsymbol{p'} \preceq n\boldsymbol{p}$ means that for all i, $n'p'_i \leq np_i$, i.e. the number of rows of class i in the subset of sd must be no larger than the number of rows of class i in sd.

Please note that the above definition does only require that there is a *subset* of the extension of *sd* on *DB* with quality $oe(\boldsymbol{p}, n, \boldsymbol{p_0}, N)$ – it does not require that there actually is a *subgroup description* with that quality. That is, the definition considers every subset of rows that is consistent with the restrictions provided by the parameters $\boldsymbol{p}$ and n. It is obvious that if an optimistic estimate for q is *tight*, then there is no optimistic estimate for q that is more conservative.

3.2   A Tight Estimate for Piatetsky-Shapiro

We will now apply our definition of tightness to the optimistic estimate published in [Wro97] for the Piatetsky-Shapiro function, and show that it is not tight. Thereafter, we will present a tight optimistic estimate for that quality function. First, we remark that in this two-class case we use p resp. p_0 to refer to the first component of $\boldsymbol{p}$ resp. $\boldsymbol{p_0}$.

Lemma 1. *The optimistic estimate $n(1 - p_0)$ presented in [Wro97] for the Piatetsky-Shapiro function $n(p - p_0)$ is not tight. The optimistic estimate $oe_{ps}^* := np(1 - p_0)$ is tight.*

Proof. We first show that oe_{ps}^* is tight. Suppose we are given a database DB and an arbitrary subgroup extension s with probability p and size n. The case $p = 0$ is trivial, so let $p > 0$; The subgroup extension s contains np rows of the first class. These rows are a subset of s with size np and a class distribution of $p = 1$; thus, this subset has quality $np(1 - p_0)$. This construction did not make any assumptions on DB or s, thus there is always a subset s' with $q(s') = oe_{ps}^*(s)$, and oe_{ps}^* is tight.

Finally, to see that $n(1 - p_0)$ is not tight it is sufficient to note that $np(1 - p_0) < n(1 - p_0)$ for some $n > 0$, $p_0 < 1$ and $p < 1$. □

3.3 Tight Estimates for the Multi-class Quality Functions Split, Gini and Pearson's χ^2

Next, we turn to the multi-class quality functions. First, please note that all multi-class quality functions $q(\boldsymbol{p}, n, \boldsymbol{p_0}, N)$ considered can be reformulated as functions $q(\boldsymbol{m}, \boldsymbol{p_0}, N)$, where $\boldsymbol{m} = (m_1, ..., m_c)$ is a vector whose components are the numbers of rows of the different classes $1, ..., c$. The m_i's can be obtained from $\boldsymbol{p}, n$ by taking the scalar product $n \cdot (p_1, ..., p_c)^T$, while $\boldsymbol{p}$ and n can be obtained from the m_i's by calculating $n = \sum_j m_j$ and $\boldsymbol{p} = \frac{1}{n} \cdot \boldsymbol{m}$. Using this new representation, we can now present a scheme of tight optimistic estimates for the multi-class quality functions in Table 1:

Lemma 2. *The following is a tight optimistic estimate for every multi-class quality function q in Table 1 (in fact, for any quality function that is convex in $\boldsymbol{p}$ resp. $\boldsymbol{m}$):*

$$oe_q^*(p_1, ..., p_c, n, \boldsymbol{p_0}, N) := \max_{m_1', ..., m_c' | m_i' \in \{0, np_i\}} \{q((m_1', ..., m_c')^T, \boldsymbol{p_0}, N)\} \quad (1)$$

The above is thus the maximum over the 2^c possible combinations of values for $m_1', ..., m_c'$, resulting from the constraint that every m_i' can either take the value 0 or np_i. Please note that although not specified in Equation 1, the case $m_i' = 0$ for every i needs not be considered.

Proof. The proof that oe_q^* is an optimistic estimate is based on the fact that all the multi-class quality functions considered in Table 1 are convex in $\boldsymbol{p}$ resp. $\boldsymbol{m}$.

First, we use the fact that by definition a tight optimistic estimate for the refinements of a subgroup s is the maximum over the quality of every possible subset of subgroup s. The subgroup s has np_1 rows of class 1, np_2 rows of class 2, and so on. Thus, every refinement s' of s consists of at most np_i rows of class i. Hence, a tight optimistic estimate for a quality function q can be calculated as follows:

$$\max_{m_1', ..., m_c' | \forall i. m_i' \in \mathbb{N}_+ \wedge 0 \le m_i' \le np_i} \{q((m_1', ..., m_c')^T, \boldsymbol{p_0}, N)\} \quad (2)$$

Please note that unlike in Equation 1, the above considers the maximum over *every* subset of rows of the subgroup s. The above expression is not only an optimistic estimate, but it is tight, which follows directly from our definition of tightness from Section 3.1.

It remains to show that Equations 1 and 2 are equivalent. This can be shown using the fact that every quality function in Table 1 is convex in its parameters $\boldsymbol{m}$. In fact, for every c-dimensional convex function f the maximum over a polyhedron $P = [0, m_1] \times [0, m_2] \times \ldots \times [0, m_c]$ is an extreme point of P, also called a vertex [Bre96, BV04]. Thus, the maximum over P can be calculated by taking the maximum of the values at every extreme point. The proof is completed by the fact that every quality function in Table 1 is convex, as shown in Appendix A. Please note that although in the Appendix we consider the extension of the quality functions to the real numbers, the extreme points are nevertheless tuples of positive natural numbers.   □

Let us now consider the computational complexity for the calculation of the tight optimistic estimate. oe_q^* involves taking the maximum of 2^c values of q, where c is the number of classes. All multi-class quality functions we considered can be reformulated to the form $q = \phi_1 + \ldots + \phi_c$ (for the Split quality function, the expressions ϕ_i summed up are $n(p_i - p_{0_i})^2$, for Gini they are $\frac{n}{N-n}(p_i - p_{0_i})^2$ and for Pearson's they are $\frac{n}{p_{0_i}}(p_i - p_{0_i})^2$). The calculation of such an expression ϕ_i involves only a constant number of subtractions and multiplications, thus the tight optimistic estimates oe_q^* have a computational complexity of $O(c\,2^c)$ primitive (add/multiply) operations. Please note that the computational complexity of oe_q^* does not depend on the size of dataset, but only on the parameters $\boldsymbol{p}, n, \boldsymbol{p_0}$ and N, which have to be calculated anyway to compute the quality of the subgroup.

3.4   A Family of Increasingly Conservative Optimistic Estimates

For large numbers of classes, c, the computational complexity, $O(c\,2^c)$, of the tight optimistic estimate oe_q^* can become problematic, as will be confirmed by the experiments in Section 4.

In this section, we present a scheme of optimistic estimates that are not tight, but faster to calculate. The estimates are increasingly conservative and at the same time increasingly complex to calculate. The idea is not to consider all 2^c combinations as done in oe_q^*, but instead to consider only combinations for d classes at a time. That is, we consider 2^d different combinations for the d selected classes; for the other classes, we only consider the two extreme cases where either *every* $m_i' = m_i$ or every $m_i' = 0$ (for classes i not within the d classes). Finally, the sum over all these maximums is calculated and used as an estimate.[3]

Here is the definition of the scheme oe_q^d of optimistic estimates, where d determines the number of classes considered at a time:

---

[3] This scheme is a generalization of the optimistic estimate proposed in [GRSW08], which considers just one class at a time (instead of d).

$$oe_q^d(\boldsymbol{p}, n, \boldsymbol{p_0}, N) :=$$

$$\sum_{j=1,d+1,2d+1,\ldots} \left[\max_{m'_j,\ldots,m'_{j+d-1}|m'_j\in\{0,np_j\}} \left(\max\Big\{ \sum_{i=j}^{j+d-1} \phi_i(\boldsymbol{m'_-},\boldsymbol{p_0},N), \right. \right. \qquad (3)$$

$$\left. \left. \sum_{i=j}^{j+d-1} \phi_i(\boldsymbol{m'_+},\boldsymbol{p_0},N) \Big\} \right) \right]$$

Here, the ϕ_i are summands of the quality functions as discussed in the previous section. The $\boldsymbol{m'_-}$ stands for the vector $(0,0,\ldots,m'_j,\ldots,m'_{j+d-1},0,\ldots,0)^T$ and $\boldsymbol{m'_+}$ for the vector $(np_0, np_1, \ldots, np_{j-1}, m'_j, \ldots, m'_{j+d-1}, np_{j+d}, \ldots, np_c)^T$; intuitively $\boldsymbol{m'_-}$ stands for the case where the subset of s includes none of the rows of the classes $1, \ldots, j-1, j+d, \ldots, c$ of the subgroup s, while $\boldsymbol{m'_+}$ stands for the case where every row of these classes is present. Of course, if c is not a multiple of d, the last summands might involve less than d classes.

Proof. Similar to the proof for oe_q^*, the proof is based on the fact that every quality function considered is convex, as shown in the Appendix. Additionally, we use the fact all quality functions considered can be brought to the form $\phi_1(\boldsymbol{m}, \boldsymbol{p_0}, N) + \ldots + \phi_c(\boldsymbol{m}, \boldsymbol{p_0}, N)$, with ϕ_i being $(\sum_j m_j)(\frac{m_i}{\sum_j m_j} - p_{0i})^2$ for Split, $\frac{(\sum_j m_j)}{N-(\sum_j m_j)}(\frac{m_i}{\sum_j m_j} - p_{0i})^2$ for Gini and $\frac{(\sum_j m_j)}{p_{0i}}(\frac{m_i}{\sum_j m_j} - p_{0i})^2$ for Pearson's. Now the following holds:

$$\max_{m'_1,\ldots,m'_c|\forall i.m'_i\in N_+\wedge 0\leq m'_i\leq np_i} \Big\{ \sum_{i=1}^c \phi_i(\boldsymbol{m'},\boldsymbol{p_0},N) \Big\} =$$

$$\max_{m'_1,\ldots,m'_c|\forall i.m'_i\in\{0,np_i\}} \Big\{ \sum_{i=1}^c \phi_i(\boldsymbol{m'},\boldsymbol{p_0},N) \Big\} \leq$$

$$\sum_{j=1,d+1,2d+1,\ldots,c} \max_{m'_1,\ldots,m'_c|\forall i.m'_i\in\{0,np_i\}} \Big\{ \sum_{i=j}^{j+d-1} \phi_i(\boldsymbol{m'},\boldsymbol{p_0},N) \Big\} \leq$$

$$\sum_{j=1,d+1,\ldots,c} \max_{m'_j,\ldots,m'_{j+d-1}|m'_j\in\{0,np_j\}} \left[\max_{m'_1,\ldots m'_{j-1},m_{j+d},\ldots m'_c|m'_i\in\{0,np_i\}} \Big\{ \sum_{i=j}^{j+d-1} \ldots\Big\} \right]$$

where ... stands for $\phi_i(\boldsymbol{m'}, \boldsymbol{p_0}, N)$. Now we make use of the fact that ϕ_i is not only convex in m_i and $m_j, j \neq i$, but also in the sum of m_j's with $j \neq i$. That is, for every set of indexes $J = \{j_1, \ldots j_n\}$ not including i, ϕ_i is convex in $\sum_{k\in J} m_k$. This follows from the fact that in all definitions of ϕ_i (i.e. for the definition for Split, Gini and χ^2) the m_j with $j \neq i$ only occur in the sum $\sum_j m_j$. Thus, for any set of indexes J that does not include i, ϕ_i could be considered as a function of the new parameter $(\sum_{k\in J} m_k)$ and the remaining $m_{k'}$, i.e. those with index $k' \notin J$. The resulting function is of the same form as ϕ_i (except that it takes less parameters) and thus is convex in $\sum_{k\in J} m_k$.

In particular, any ϕ_i with $j \leq i \leq j+d-1$ is convex in $\sum_{k \in \{0,\ldots,j-1,j+d,\ldots c\}} m_k$. Therefore, it is sufficient to consider the case where the sum is minimal or maximal, that is when either all these m'_k's are zero or all have value np_k. So the above is bounded by

$$\sum_{j=1,d+1,\ldots,c} \left[\max_{m'_j,\ldots,m'_{j+d-1} \mid m'_j \in \{0,np_j\}} \left(\max_{(\forall k.\, m'_k=0),(\forall k.\, m'_k=np_k)} \sum_{i=j}^{j+d-1} \phi_i(\boldsymbol{m'}, \boldsymbol{p_0}, N) \right) \right]$$

This is equivalent to oe_q^d and the proof is completed. $\square$

Some Considerations on the new Optimistic Estimates. We will now consider some properties of the optimistic estimates oe_q^d. As before, we use c to denote the number of classes. The computation of oe_q^d involves the evaluation of $O(\frac{c}{d} 2^d d) = O(c2^d)$ different expressions ϕ_i, meaning that the higher the number d of classes considered at a time, the higher the computational cost. We remark that in an implementation of the function oe_q^d, only those classes where $m_i > 0$ have to be considered, meaning that $O(c2^d)$ is only the worst case.

Lemma 3. oe_q^{2d} *is at least as conservative as* oe_q^d. oe_q^d *is tight if either* $d \geq c$, *or* $c = 2$ *and* $d \geq 1$.

Proof. It is easy to see that if $c \leq d$, then $oe_q^d = oe_q^*$, because the number of classes d in oe_q^d for which every combination is considered is exactly c, as in oe_q^*. oe_q^{2d} is at least as conservative as oe_q^d because oe_q^{2d} considers every combination of classes considered by oe_q^d (and some more combinations).

In the special case $c = 2$, oe_q^1 is tight because the set of indexes $\{0,\ldots,j-1,j+d,\ldots c\}$ considered in the second sum of Equation 3 actually involves only one class index, and hence effectively every combination of the two classes is considered, just as in oe_q^*. To show that oe_q^d is not tight otherwise, it is sufficient to have one example. The experiments in Section 4 have plenty of them, and the next paragraph also presents one. Here is another one: consider $\boldsymbol{p_0} = (0.1, 0.45, 0.45)$ (for $c = 3$) respectively $\boldsymbol{p_0} = (0.1, 0.3, \frac{0.3}{c-3}, \ldots, \frac{0.3}{c-3}, 0.3)$ (for $c > 3$). Furthermore, consider a subgroup s with $\boldsymbol{m} = (10, 10, 0, 0, \ldots 0, 10)$. It is easy to verify that of all subsets of $\boldsymbol{m}$, $\boldsymbol{m'} = (10, 0, 0, \ldots 0, 0)$ has the highest quality. However, the last summand of oe_q^d would consider the values 10 and 0 for m'_c only in combination with *either* both $m'_1 = 10$ and $m'_2 = 10$ *or* both $m'_1 = 0$ and $m'_2 = 0$. That is, it would not consider the actual maximizing combination $(10, 0, 0, \ldots 0, 0)$, but instead will provide an overoptimistic estimate. $\square$

An Example. To illustrate the effect of different optimistic estimates, let us reconsider the example from Figure 1. In particular, let us consider the subgroup description $\{Region = Urban\}$. The corresponding subgroup consists of the first three rows of the dataset and has a probability distribution of $p_{High} = 1/3$, $p_{Med} = 2/3$, and $p_{Low} = 0$.

For the Split quality function, we get a tight optimistic estimate of 0.176 for the quality of the refinements of $\{Region = Urban\}$. Using the suboptimal

estimate oe_q^1, we only get the estimate 0.255. While both estimates can be used by a subgroup discovery algorithm, if the minimal required quality is 0.195 only the tight estimate of 0.176 would allow to prune all subgroups description below $\{Region = Urban\}$ (In fact, 0.195 is exactly the minimal required quality at that point, if the algorithm sketched in Section 4 is used).

4   Experimental Evaluation

Sketch of a Subgroup Discovery Algorithm. To evaluate the impact of different optimistic estimates, we used a branch and bound depth-first-search (DFS) algorithm similar to OPUS^O [Web95] (without optimistic reordering). The optimistic estimates are used to prune as large a part of the search space as possible, and furthermore determine the order in which the nodes are expanded during DFS. Our implementation makes use of FP-Trees [HPY00] to speedup the calculation of the parameters $\boldsymbol{p}$ and n of a subgroup, as first proposed in [AP06]. The overall algorithm, called DpSubgroup, is described in more detail in [GRSW08].

Datasets and Results. We evaluated the impact of the different optimistic estimates on five datasets: four datasets from the UCI Machine Learning Repository
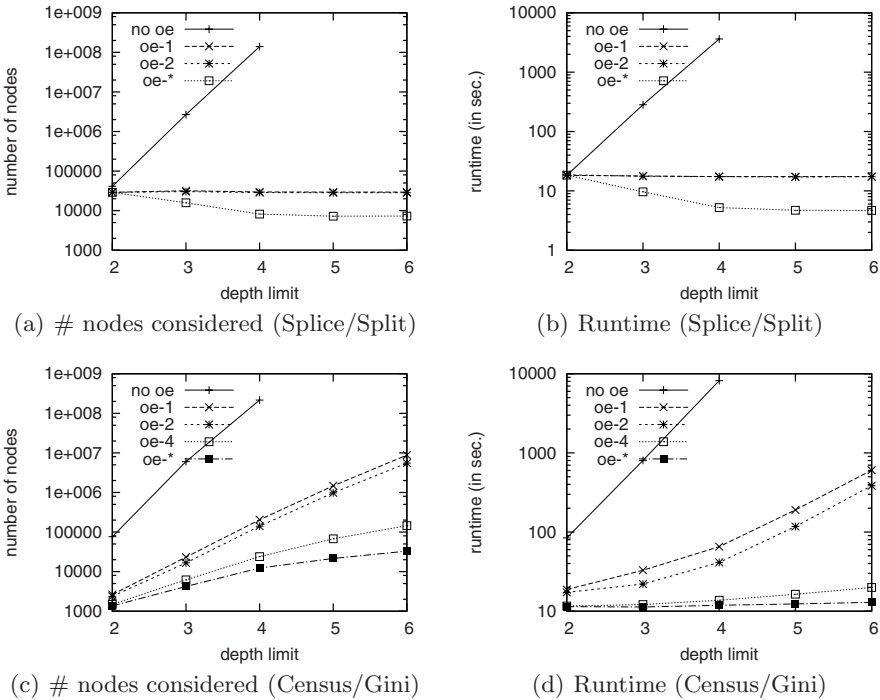

(a) # nodes considered (Splice/Split)   (b) Runtime (Splice/Split)

(c) # nodes considered (Census/Gini)   (d) Runtime (Census/Gini)

Fig. 2. Results for the Splice and Census dataset (with 3 resp. 5 classes) using the Split resp. the Gini quality function. The curves show the number of nodes considered (left images) resp. the runtime (right images) for different optimistic estimates.

[AN07] and one real-world dataset. In particular, we used the "Mushroom" dataset with 8124 rows, 22 attributes and 2 classes; the "Soybean" dataset with 683 rows, 35 attributes and 19 classes; the "splice" dataset with 3190 rows, 62 attributes and 3 classes; a sample of 30.000 rows of the UsCensus1990 database with 68 attributes and 5 classes (we used "dIncome1" as class attribute); and finally a prescription dataset with 29 attributes, 60488 rows and 11 classes from the *iWebCare* project (http://iwebcare.iisa-innov.com/). In all experiments, we searched for the top 100 groups on an Intel Core 2 Duo T7500 with 2 GB of RAM under Windows XP.

Figure 2 shows the performance results on the datasets Census and Splice, using different quality functions. The horizontal axis shows the depth limits for the subgroup discovery, that is a maximum length of the subgroup descriptions considered. The different curves show the results using different optimistic estimates: "oe-d" stands for subgroup discovery with oe_q^d, "oe-*" shows the results using the tight optimistic estimate oe_q^*, and finally "no oe" shows the performance without any optimistic estimate pruning. It is worthwhile to note that "no oe" essentially corresponds to the algorithm SD-Map (which also makes use of FP-Trees but does not use optimistic estimate pruning), because this algorithms has been shown to outperform all other exhaustive algorithms like Apriori-SD [KLJ03] by an order of magnitude [AP06].

The figure shows both the number of nodes explored during the subgroup discovery and the overall runtime, using a logarithmic scale. As expected, the number of nodes considered depends on the optimistic estimate used: The higher the degree d in oe_q^d, the less nodes are considered. The use of the tight optimistic estimate oe_q^* results in a minimal number of considered nodes.[4] Similar to the number of nodes, the runtime is affected by the optimistic estimate used. Although the performance ratio does not exactly correspond to the node ratio, the ordering of the optimistic estimates is the same. The performance gain using pruning can become as large as an order of magnitude and more.
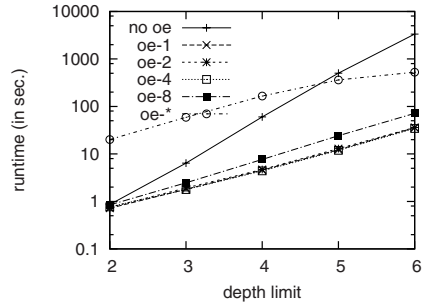
Figure 3 shows the results for the Soybean and the Mushroom dataset. We first consider the results for the Soybean dataset (subfigures (a)-(c)): Again, the higher the degree d of the optimistic estimate oe_q^d, the less nodes are considered, with oe_q^* being optimal (Figure 3(a)). Regarding the runtime, however, the situation is different: Figure 3(b) shows that the runtime is minimal when a pruning level d is between 1 and 4. The reason is that although the more conservative optimistic estimates reduce the number of nodes considered, their calculation is more expensive. This effect becomes apparent in the Soybean dataset because it has much more classes than the earlier datasets.

The above experiment shows that for datasets with a large number of classes, it can be appropriate to use a non-tight optimistic estimate. However, the best pruning level depends on the *ratio* of the costs for the calculation of the optimistic
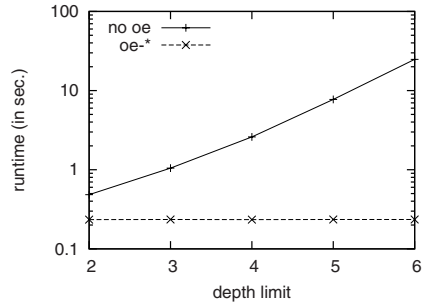
---

[4] It is interesting to note that the number of nodes and the runtime sometimes *decreases* when the depth limit increases. The reason is that the algorithm quickly finds subgroups with a very high quality at a higher level, which allows to prune a larger part of the search space than possible if only shorter subgroups were considered.
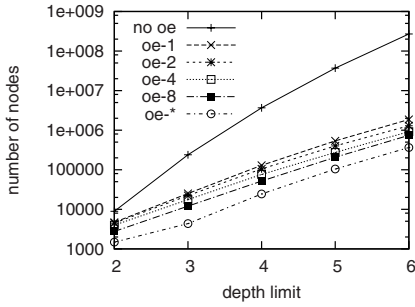
(a) # nodes considered (Soybean/Gini)   (b) Runtime using FpTree (Soybean/Gini)
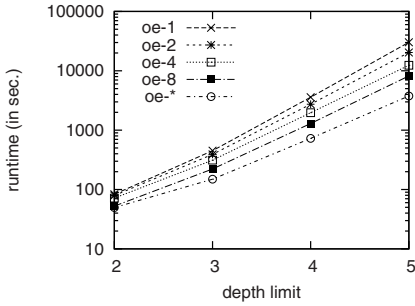
(c) Runtime using *persistent* FpTree   (d) Runtime (Mushroom/P.-S.)

Fig. 3. Number of nodes (a) and runtime (b, c) on the Soybean dataset (using the Gini quality function), and runtime on the mushroom dataset (d) (using the Piatetstky-Shapiro quality function)

estimate, and the cost for the calculation of the parameters of a subgroup. So far we used FP-Trees, which allowed a very fast calculation of n and $\boldsymbol{p}$. However, if the dataset is very large it might not be possible to keep the FP-Tree in main memory. To see how the situation changes if the calculation of the parameters n and p becomes more expensive, we have run some experiments using a *persistent* FP-Tree, i.e. an implementation where the FP-Trees are stored on disk [HPYM04]. Our prototypical implementation is based on the object database db4o [PEHH06]. Figure 3(c) shows the resulting runtime: In this setting, the use of conservative optimistic estimates pays off again, with oe_q^* resulting in the fastest calculation (the number of nodes is, of course, unaffected).

Finally, Figure 3(d) shows the runtime for the Mushroom dataset, using the Piatetstky-Shapiro quality. As in the multi-class case, the use of optimistic estimates results in a significant speedup in this two-class example. The results from the prescriptions dataset do not significantly differ, so we merely used them in the summarizing table in the next paragraph.

Summary. The above results show that the optimistic estimates presented in this paper have a significant impact on the performance of the subgroup discovery. The use of the (fastest to calculate) optimistic estimates oe_q^1 and oe_{ps}^* never

slows down the execution but instead results in a tremendous speedup. The performance gain reaches an order of magnitude at relative small depth bounds (about 2 to 4) and gets even larger when the depth limit continues to increase.

The use of more conservative optimistic estimates allow to further speedup the subgroup discovery. The following table summarizes the performance gain over oe_q^1 achieved using the optimistic estimates oe_q^d and oe_q^* (with $d > 1$). For different values of d (i.e. 2, 3, 4 and ∞), it compared the performance with that achieved using oe_q^1, aggregated over different quality functions and datasets. In particular, it shows the relative runtime (the runtime using the more conservative optimistic estimate, divided by the runtime using oe_q^1) in the best experiment (Minimal), the worst experiment (Maximal) and on average. The table is based on a total of 248 experiments.

	oe_q^2	oe_q^3	oe_q^4	oe_q^*
Minimal relative runtime compared to oe_q^1	62%	21%	3%	1%
Average relative runtime compared to oe_q^1	93%	69%	63%	847%
Maximal relative runtime compared to oe_q^1	113%	118%	135%	3640%

The table shows that the larger d, the more the runtime can decrease (due to the stronger pruning) - but it can also increase (due to the computation time of the optimistic estimate). Overall, the use of oe_q^4 (which is tight if the number of classes is ≤ 4) is a safe choice in most situations. While in the worst example, it was slower by 35% than oe_q^1, on average it was faster, taking only 63% of oe_q^1's runtime; In the best example, it even reduced the runtime to 3% of that of oe_q^1.

5   Summary and Discussion

In this paper, we have pursued the investigation of optimistic estimates for fast subgroup discovery, started in [Wro97]. In particular, we considered and formalized the concept of *tight* optimistic estimates. We have shown that the optimistic estimate proposed in [Wro97] is not tight, and have presented new tight optimistic estimates, including tight optimistic estimates for several multi-class quality functions.

While the use of tight optimistic estimates minimizes the number of subgroups considered, their calculation sometimes becomes quite time consuming. To cope with this difficulty, we have presented a family of increasingly efficient approximations of the tight optimistic estimates. While these estimates are (in general) not tight, they allow to trade more conservative estimates for faster computation and thereby provide a mean to select an optimistic estimate with the right ratio of conservative-ness and computational cost.

The results are interesting both from a theoretic and a practical point of view. On the theoretical side, the notions of conservative and tight optimistic estimates allow to compare optimistic estimates. On the practical side, our experiments show that the use of the new optimistic estimates oe_{ps}^*, oe_q^d and oe_q^* result in a significant speedup compared to current state-of-the-art algorithms like SD-Map [AP06]. While for problems with a relatively small number of classes oe_q^* is the estimate of choice, for datasets with a larger number of classes (more than 6 or

so) the optimal choice depends on more factors. Overall, the use of oe_q^4 is a safe choice in most situations.

The idea to perform pruning based on an optimistic evaluation of the search space below a node was already investigated before the introduction of optimistic estimate into subgroup discovery. In particular, Webb [Web95] considers optimistic pruning for rule learning tasks in his OPUS search algorithm. The concept is also applied in other pattern-mining tasks involving non-monotonic objective functions, like tiling databases [GGM04]. Of course, the objective function considered in tasks like tiling are different than in subgroup discovery.

In future work, we plan to investigate optimistic estimates for other quality function, and to extend the concept of tight optimistic estimates to numeric target attributes. We would also like to investigate whether the optimistic estimates oe_q^d can be improved by some kind of heuristic grouping of the classes, with the idea not just to build arbitrary groups of d attributes. It would also be interesting to combine the optimistic estimates presented in this paper with other approaches to subgroup discovery, in particular with sampling-based approaches [SW00]. Altogether, we believe that the definition of tight optimistic estimates and the new optimistic estimates presented in this paper can be a valuable instrument in a wide range of subgroup discovery algorithms.

References

[ABP06]   Atzmüller, M., Baumeister, J., Puppe, F.: Introspective subgroup analysis for interactive knowledge refinement. In: Sutcliffe, G., Goebel, R. (eds.) FLAIRS Conference, pp. 402–407. AAAI Press, Menlo Park (2006)

[AN07]   Asuncion, A., Newman, D.J.: UCI machine learning repository (2007)

[AP06]   Atzmüller, M., Puppe, F.: SD-map - a fast algorithm for exhaustive subgroup discovery. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 6–17. Springer, Heidelberg (2006)

[BFOS84]   Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (eds.): Classification and Regression Trees. Wadsworth (1984)

[Bre96]   Breiman, L.: Technical note: Some properties of splitting criteria. Machine Learning 24(1), 41–47 (1996)

[BV04]   Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)

[GGM04]   Geerts, F., Goethals, B., Mielikäinen, T.: Tiling databases. In: Suzuki, E., Arikawa, S. (eds.) DS 2004. LNCS (LNAI), vol. 3245, pp. 278–289. Springer, Heidelberg (2004)

[GRSW08]   Grosskreutz, H., Rüping, S., Shaabani, N., Wrobel, S.: Optimistic estimate pruning strategies for fast exhaustive subgroup discovery. Technical report, Fraunhofer Institute IAIS (2008), `http://publica.fraunhofer.de/eprints/urn:nbn:de:0011-n-723406.pdf`

[HPY00]   Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Chen, W., Naughton, J.F., Bernstein, P.A. (eds.) SIGMOD Conference, pp. 1–12. ACM, New York (2000)

[HPYM04] Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Min. Knowl. Discov. 8(1), 53–87 (2004)

[KLGK07] Kralj, P., Lavrac, N., Gamberger, D., Krstacic, A.: Contrast set mining through subgroup discovery applied to brain ischaemina data. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 579–586. Springer, Heidelberg (2007)

[KLJ03]   Kavsek, B., Lavrac, N., Jovanoski, V.: Apriori-sd: Adapting association rule learning to subgroup discovery. In: R. Berthold, M., Lenz, H.-J., Bradley, E., Kruse, R., Borgelt, C. (eds.) IDA 2003. LNCS, vol. 2810, pp. 230–241. Springer, Heidelberg (2003)

[Klö96]   Klösgen, W.: Explora: A multipattern and multistrategy discovery assistant. In: Advances in Knowledge Discovery and Data Mining, pp. 249–271 (1996)

[Klö02]   Klösgen, W.: Subgroup Discovery. In: Handbook of Data Mining and Knowledge. Oxford University Press, Oxford (2002)

[KM02]   Klösgen, W., May, M.: Spatial subgroup mining integrated in an object-relational spatial database. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) PKDD 2002. LNCS (LNAI), vol. 2431, pp. 275–286. Springer, Heidelberg (2002)

[LCGF04] Lavrac, N., Cestnik, B., Gamberger, D., Flach, P.A.: Decision support through subgroup discovery: Three case studies and the lessons learned. Machine Learning 57(1-2), 115–143 (2004)

[LKFT04] Lavrac, N., Kavsek, B., Flach, P., Todorovski, L.: Subgroup discovery with cn2-sd. Journal of Machine Learning Research 5, 153–188 (2004)

[PEHH06] Paterson, J., Edlich, S., Hörning, H., Hörning, R.: The Definitive Guide to db4o. Apress, Berkely (2006)

[SW00]   Scheffer, T., Wrobel, S.: A sequential sampling algorithm for a general class of utility criteria, pp. 330–334. ACM Press, New York (2000)

[Web95]   Webb, G.I.: Opus: An efficient admissible algorithm for unordered search. J. Artif. Intell. Res (JAIR) 3, 431–465 (1995)

[Wro97]   Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Komorowski, J., Żytkow, J.M. (eds.) PKDD 1997. LNCS, vol. 1263, pp. 78–87. Springer, Heidelberg (1997)

A   Proof of Convexity of the Multi-class Quality Functions

A.1   Split and Pearson's χ^2

Both the Split and Pearson's χ^2 quality functions are nonnegative weighted sums. The nonnegative weighted sum of convex functions is convex ([BV04]), hence it is sufficient to show that the summands are convex. The summand of both quality functions can be brought to the following form

$$\phi_i = c_i(\sum_j n_j)(\frac{n_i}{\sum_j n_j} - p_{0_i})^2$$

where in the case if Split $c_i = 1$ and in the case of Pearson's $c_i = \frac{1}{p_{0_i}}$.

We only need to consider the case where $c_i = 1$, because the c_i's are merely positive weights. The ϕ_i's are twice differentiable, thus it is sufficient to show that the *Hessian* or second derivative is positive semidefinite [BV04]. We only consider the first summand ϕ_1, as the other cases are analog. It is somewhat laborious but straightforward to verify that

$$\nabla^2 \phi_1 = \frac{2}{(\sum_j n_j)^3} \begin{bmatrix} (\sum_{j\neq 1} n_j)^2 & -n_1(\sum_{j\neq 1} n_j) & \dots & -n_1(\sum_{j\neq 1} n_j) \\ -n_1(\sum_{j\neq 1} n_j) & (n_1^2) & \dots & (n_1^2) \\ \dots & \dots & \dots & \dots \\ -n_1(\sum_{j\neq 1} n_j) & (n_1^2) & \dots & (n_1^2) \end{bmatrix}$$

The above matrix can be brought to the following form

$$GG^T$$

with $G = (\sum_{j\neq 1} n_j, -n_1, ..., -n_1)^T$, that is there is a Cholesky decomposition and hence the matrix is positive-definite [BV04].

A.2   Gini

The fact that the Gini quality function is convex can be derived from previous work in the area of decision tree construction. This comes from the fact that the Gini quality functions is based on the *Gini index*, used as a splitting criterion in the construction of decision trees [Bre96, BFOS84]. The Gini index is defined as $G(p) = \sum_j p_j(1 - p_j)$ and measures the "impurity" of a distribution p. The gain in impurity resulting from a split is defined as $\Theta(s) = G(\boldsymbol{p_0}) - P_L G(\boldsymbol{p_l}) - P_R G(\boldsymbol{p_r})$ and was used as a *goodness of a split* measure. Here, $\boldsymbol{p_0}$ denotes the distribution over the classes in the overall population, $\boldsymbol{p_l}$ and $\boldsymbol{p_r}$ denote the distribution in the left and the right subpopulation resulting from a split, P_L denotes the proportion of the population send to the left by the split and $P_R = 1 - P_L$ denotes the proportion send to the right.

This goodness of split was turned into the Gini quality function [Klö96]:

$$G(\boldsymbol{p_0}) - g * G(\boldsymbol{p}) - (1 - g) * G(\boldsymbol{p^*})$$

where $\boldsymbol{p_0}$ is (as before) the class distribution in the overall population, $\boldsymbol{p}$ the class distribution in the subgroup, $\boldsymbol{p^*}$ the probability distribution in the remainder, i.e. in the examples from the overall population not in the subgroup, and finally g the generality of the subgroup. It is easy to verify that the components of p^* are defined by $p_i^* = \frac{Np_{0i} - Ngp_i}{N(1-g)} = \frac{p_{0i} - gp_i}{1-g}$. By inserting this definition of p^* we obtain the following, more familiar-looking definition of the Gini quality,

$$\frac{g}{1-g} \sum_j (p_i - p_{0i})^2 = \frac{n}{N-n} \sum_j (p_i - p_{0i})^2$$

[Bre96] shows that $g * G(\boldsymbol{p}) + (1 - g) * G(\boldsymbol{p^*})$ is concave in the proportion of the probabilities that are sent to the left (which he calls α). We remark that Breiman

uses the term convex in the meaning of "convex downward"; We use the term convex in the opposite sense, as [BV04] (Breimans notation also differs from ours in other respects). Now our vector $\boldsymbol{m}$ can be obtained from α by an affine mapping, namely $m_i = N p_{0_i} \alpha_i$, which implies that $g * G(\boldsymbol{p}) + (1 - g) * G(\boldsymbol{p^*})$ is also concave in $\boldsymbol{m}$, because affine mappings preserves concavity. Hence the Gini quality function is convex in $\boldsymbol{m}$.