International Journal of Approximate Reasoning 80 (2017) 199–213

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

# Bounds on skyline probability for databases with uncertain preferences

Arun K. Pujari, Vineet Padmanabhan, Venkateswara Rao Kagita *

*Artificial Intelligence Lab, School of Computer & Information Sciences, University of Hyderabad, Hyderbad 500046, AndhraPradesh, India*

ARTICLE INFO

*Article history:*
Received 18 April 2015
Received in revised form 11 September 2016
Accepted 13 September 2016
Available online 20 September 2016

*Keywords:*
Bounds
Skyline probability
Uncertain preferences

ABSTRACT

For determining skyline objects for an uncertain database with uncertain preferences, it is necessary to compute the skyline probability of a given object with respect to other objects. The problem boils down to computing the probability of the union of events from the probabilities of all possible joint probabilities. Linear Bonferroni bound is concerned with computing the bounds on the probability of the union of events with partial information. We use this technique to estimate the skyline probability of an object and propose a polynomial-time algorithm for computing sharp upper bound. We show that the use of partial information does not affect the quality of solution but helps in improving the efficiency. We formulate the problem as a Linear Programming Problem (LPP) and characterize a set of feasible points that is believed to contain all extreme points of the LPP. The maximization of the objective function over this set of points is equivalent to a bi-polar quadratic optimization problem. We use a spectral relaxation technique to solve the bi-polar quadratic optimization problem. The proposed algorithm is of $O(n^3)$ time complexity and is the first ever polynomial-time algorithm to determine skyline probability. We show that the bounds computed by our proposed algorithm determine almost the same set of skyline objects as that with the deterministic algorithm. Experimental results are presented to corroborate this claim.



## 1. Introduction

With the rapid increase of uncertain data, query processing on uncertain databases has become an important research topic in the database community. For a given dataset, skyline computation aims at retrieving the set of all skyline objects. A skyline object is an object that is not *dominated* by any other object. An object $p$ *dominates* an object $q$ if $p$ is strictly better than $q$ on at least one dimension and $p$ is better than or equal to $q$ on the remaining dimensions. We assume that the domains of dimensions are ordered. Reasoning and learning with preferences, particularly uncertain preferences, is a major area of investigation in AI. Databases having objects with non-quantifiable attributes can only be compared in terms of ranking or preferences. While annotating or eliciting preferences from the users, it becomes easy to get pairwise preferences of attribute values instead of ranking all the attribute values. Thompson et al. [1] argue that obtaining relative preference judgment for a pair of items is easier and less prone to error and enables fine-grained comparison. Pairwise assessment is also valuable for inferring global ranking. Such annotation through crowdsourcing may result in inconsistent

* Corresponding author.
*E-mail addresses:* akpcs@uohyd.ernet.in (A.K. Pujari), vineetcs@uohyd.ernet.in (V. Padmanabhan), venkateswar.rao.kagita@gmail.com (V.R. Kagita).

http://dx.doi.org/10.1016/j.ijar.2016.09.004
0888-613X/

preferences such as $a$ is preferred to $b$ by one annotator and $b$ is preferred to $a$ by another. In such cases, some probability measure can be assigned to pairwise preferences.

In this paper, we address the problem of identifying objects which are of interest in the Pareto sense with pairwise preference probability of attribute values of a set of given multi-attribute objects. This problem has got several real-life applications in AI and Machine Learning. In aesthetic photo ranking problem [2], digital photos have high-level features such as *rule of thirds, saliency, color* and *balance*. These are non-quantifiable features, and it becomes relevant to have pairwise preferences among these features. In other words, an assessor can compare the photos to say which one is better than the other in terms of color or balance or any other attribute value. Let us consider another situation to emphasize that such pairwise preferences can also be uncertain. Consider the problem of selecting interesting TV programs. TV programs have different attributes such as *category, genre, duration*, etc. Different categories of TV programs are family-serial, reality-show, movies, etc., and it is hard to define an ordering of different categories/genres/other attributes. On the other hand, it is possible to specify, based on a record of usage, a probability that 'movies' is preferred to 'reality shows' and a probability of 'reality shows' is preferred to 'movies'. We can arrive at a preference probability of one TV program over the other as the product of attribute-wise preference probabilities. Skyline probability of a TV program is then the probability that no other program is preferred to this one on all attributes. It is of interest to identify TV programs with high skyline probability. The objective is to shortlist *skyline objects* such that no other object is preferred over these (skyline) objects on all high-level features. When the preferences are uncertain, and expressed in terms of preference probability, the problem is to determine the probability of an object being a skyline object.

Many algorithms have been proposed in the literature to deal with skyline computation over uncertain databases [3–11]. Computation of probabilistic skyline by efficiently pruning non-skyline objects was investigated by Pei et al. [3] in 2007. Kim et al. [4] proposed techniques to compute exact skyline probabilities of all objects for a given database with uncertain data. Distributed skyline queries over uncertain data was investigated in Ding et al. [5] . Yong et al. [6] studied skyline queries over uncertain data with 'maybe' confidence semantics and further extended to the scenarios wherein data dependency exists. A framework for evaluating skyline queries over uncertain dimensions was introduced in Sadd et al. [7]. The framework is suitable for performing skyline queries in uncertain autonomous databases. For determining probabilistic skyline in an uncertain data model an efficient parallel algorithm using Map-Reduce was devised in Park et al. [8]. Other related works include updating probabilistic skyline in uncertain data streams by Liu et al. [9] as well as computing probabilistic skylines against the most recent uncertain objects in a data stream by Zhang and Li et al. [10]. Exploring best probabilistic skyline query for efficient retrieval of interesting skyline points in uncertain databases was introduced in Le et al. [11]. All these work deal with probabilistic skylines over uncertain data.

The aim of this study is to compute a probability of a tuple being a skyline point; this is called skyline probability. In this setting, skyline computation is concerned with retrieving objects having skyline probability above a threshold $\tau$. Skyline computation under uncertainty can also hold when the attribute preferences are uncertain (instead of uncertainties of values). Multidimensional objects may have fixed attribute values while preferences among those values maybe uncertain. This is relevant for databases with categorical attributes where the preference ordering might be uncertain. Sacharidis et al. [12] were the first to propose a method for computing skyline probabilities of data with uncertain preferences. Later on Zhang et al. [13] made more concrete investigations of the problem and established that the problem of computing skyline probability of an object with predefined preferences is #P-complete. In our previous work, we introduced a concept of zero-contributing set [14,15] to avoid redundant computation in an exponential search space. In [15], an efficient algorithm is proposed to compute the skyline probability over uncertain preferences. In another work [16], we propose a different problem for determining skyline objects in a database with uncertain preferences for a specified threshold $\tau$ ($\tau$ is given beforehand.). There is no polynomial time method existing in the literature to compute the exact value of skyline probability when the preferences between the attribute values are uncertain.

It is worthwhile to investigate polynomial-time methods for estimating the probabilities rather than computing the exact value using apparently exponential-time technique. Since determining skyline points is essentially identifying objects with skyline probability exceeding a certain threshold, it justifies to estimate (as close as possible) skyline probability instead of investing efforts for computing the exact value. In this work, we address the problem of computing the best upper bound of skyline probability and if the upper bound is above the threshold then the object is returned as a skyline point. We propose an $O(n^3)$ algorithm to compute the upper bound on skyline probability and show that the set of skyline objects computed using this upper bound is almost the same as that of the set computed by using exact skyline probabilities.

The problem of computing bounds on skyline probability is shown to be equivalent to a linear programming problem (LPP). We identify a set of feasible solutions which contains the extreme points of linear programming (LP). The objective function of LP over the set of feasible solutions can be optimized by maximizing a quadratic function over the set of integers. The relaxed quadratic optimization is solved by a standard derivative method, and the integer approximation is obtained by the spectral method. Thus, the upper bound of the skyline probability is obtained by finding the eigenvector corresponding to the smallest eigenvalue of the covariance matrix. We propose here three different bounds on skyline probability and demonstrate empirically that the computation of all skyline points using these bounds is more accurate than the earlier polynomial method namely, sampling based Monte-Carlo estimation technique [13]. We also show that our methods outperform earlier methods in terms of the absolute error. Even when the object's skyline probability is less than the threshold, the bounds may exceed the threshold resulting in spurious (or, false) detection. However, in practice, if the bound is reasonably sharp with respect to the distribution of skyline probabilities of all query objects and the threshold,