Position Paper: Subspace Nearest Neighbor Search - Problem Statement, Approaches, and Discussion

Michael Hund[1], Michael Behrisch[1], Ines Färber[2], Michael Sedlmair[3], Tobias Schreck[4], Thomas Seidl[2], and Daniel Keim[1]

[1] University of Konstanz, Germany, firstname.lastname@uni-konstanz.de
[2] RWTH Aachen University, Germany, lastname@informatik.rwth-aachen.de
[3] University of Vienna, Austria, firstname.lastname@univie.ac.at
[4] Graz University of Technology, Austria, firstname.lastname@cgv.tugraz.at

Abstract. Computing the similarity between objects described in low- and high-dimensional feature spaces is a central task in many application scenarios, including information retrieval and data mining. For finding k-nearest neighbors to a user query, typically a ranking is computed based on a predetermined set of data dimensions and distance function, constant over all possible queries. However, many feature spaces contain a large number of dimensions, some or many of which may contain noise, irrelevant, redundant, or contradicting information. More specifically, the relevance of dimensions may depend on the query object itself, and in general, different dimension sets (subspaces) may be appropriate for a query. Approaches for feature selection, extraction and weighting typically provide a global subspace selection, which may not be best for all possibly queries. In this position paper, we frame a new research problem, called *subspace nearest neighbor search*, aiming at determining query-dependent subspaces for nearest neighbor search. We describe relevant problem characteristics, relate to existing approaches for subspace analysis and feature selection, and outline potential application benefits.

Keywords: Nearest neighbor search, subspace analysis and search, subspace clustering, subspace outlier detection

1 Introduction

Searching for similar objects is a crucial task in many applications, such as image or information retrieval, data mining, biomedical applications and e-commerce. Typically *k-nearest neighbor queries* are used to compute *one result* list of similar objects derived from a given set of data dimensions and distance function. However, the consideration of all dimensions and a single distance function may not appropriate for all queries, as we will discuss in the following.

For datasets with a high number of dimensions, similarity measures loose their discriminative ability as similarity values concentrate about their respective means. This phenomenon, known as the curse of dimensionality [2], leads to an instability of nearest neighbor queries in high-dimensional spaces. The effects of the curse of dimensionality are especially strong if the proportion of irrelevant dimensions is high.

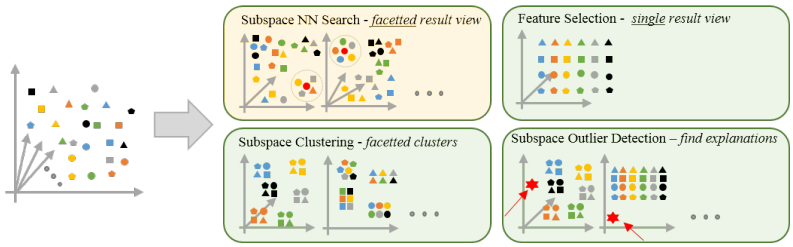


Fig. 1: Focus of Subspace Nearest Neighbor Search and related approaches.

Imagine the following example from a clinical application scenario: A physician is treating a patient with an *unknown disease* and wants to retrieve similar patients along with their medical history (treatment, outcome, etc.). In the search process the physician might, for instance, be faced with a high number of recorded profile characteristics and unrelated diseases, respectively their symptoms. If, however, the nearest neighbor ($\mathcal{NN}$) search is based on all available features, the most similar patients are probably not suited to guide the diagnostic process as irrelevant dimensions, such as the hair color, dominate the search process. Meaningful conclusions can only be drawn if the dimensions which are *characteristic* for this disease are considered. The challenging question is therefore, what is the relevant subset of dimensions (=*subspace*) specific for a certain query? Do multiple relevant subspaces for a query exist? Many other application examples can be found, where $\mathcal{NN}$ search in query-dependent subspaces is potentially relevant. To name a few: In multimedia retrieval a query may depend on the input object type; in recommender systems a query may depend on user preferences, or a kNN-classifier may depend on the class label.

Consequently, we can derive a novel research challenge, which we call *subspace nearest neighbor search*, for short $\mathcal{SNNS}$. Its central idea is to incorporate a query-dependency focus into the relevance definition of subspaces. As one example, $\mathcal{SNNS}$ allows deriving discriminative subspaces in which the $\mathcal{NN}$ of a query can be separated from the rest of the data. Alternatively, in the example from above the physician will focus on a large number of dimensions to maximize the semantic interpretability of the $\mathcal{NN}$ along with the query-dependent subspace.

$\mathcal{SNNS}$, although inspired by works in subspace clustering and -search, goes beyond these ideas by providing query-dependency as the main goal. This leads to a different problem definition. In $\mathcal{SNNS}$, our goal is to (1) detect *query dependent* and *previously unknown subspaces* that are interesting, and (2) derive the corresponding nearest neighbor set that is similar to the query within that corresponding subspace. This paper will detail on the following questions: "What is a relevant subspace for a given query object", "How can we computationally extract this relevance information?", "How can we combine ideas from subspace clustering, outlier detection, or feature selection for our purposes?"

2 Related Problems

$\mathcal{SNNS}$ relates to and draws on ideas from Feature Selection, Subspace Clustering, Subspace Outlier Detection and others. Next, we give a concise overview of these fields and related them to $\mathcal{SNNS}$. An overview is also given in Fig. 1.

Feature selection, extraction and weighting. The aim of feature selection [10] is to determine one subspace that improves a global optimization criterion for a given data set (e.g., classification error). Most algorithms do not make explicit to the user which and why features have been selected. Some visual and interactive approaches as described in [7] provide visualizations about the similarity and interestingness of dimensions which helps the user in selecting an appropriate subspace. There are two main differences to $\mathcal{SNNS}$: The derived subspaces are (1) query independent, and (2) there is only a single result view, unlike $\mathcal{SNNS}$ aiming for a *faceted result view* of multiple, independent subspaces.

Subspace Clustering. Subspace clustering aims to find high-quality clusters in different axis-parallel or arbitrarily-oriented subspaces. Main works are surveyed for example in [9]. The approaches are based on a heuristic to measure for the cluster quality. Then, specific search methods aim to find subspaces providing good clusters. The computation of clusters and subspaces is tightly coupled, but also some decoupled approaches exist, e.g. [8]. Subspace clustering differs from $\mathcal{SNNS}$ by considering a specific (clustering) application.

Subspace Outlier Detection. Works in this area fall into two groups with the aim of searching for subspaces in which (1) an arbitrary, or (2) a user-defined record is considered as outlier [13]. Like before, the process of subspace search and of outlier detection can be coupled or separated with an appropriate subspace quality criterion. Hereby, the criterion measures the degree of outlierness e.g. by item separability [11]. Subspace outlier detection is similar to $\mathcal{SNNS}$ as both approaches aim for query-dependent subspaces. However, the relevance of a subspace differs significantly. $\mathcal{SNNS}$ searches for objects that are similar to the query, while subspace outlier detection seeks for outliers.

Query-dependent Subspace Search. In [4], Hinneburg et al. propose to determine a (single) query-dependent subspace to improve $\mathcal{NN}$ queries by overcoming the challenges of the *curse of dimensionality*. In their paper they describe an approach to measure quality of a subspace by the separability between all data records and the $\mathcal{NN}$ of the query. In their evaluation they show that query-dependent subspaces reduce the error of a $\mathcal{NN}$-classification substantially. The work by Hinneburg et al. can be seen as initial approach on $\mathcal{SNNS}$ and therefore most closely related to our work. However, the general aims of [4] differs, as it does not search for different $\mathcal{NN}$ in multiple, different subspaces.

Other Related Problems. Besides these main lines, another related field is that of recommender systems [1], which focuses on similarity aspects to retrieve items of interest. Intrinsic dimensionality estimation [3] shares the intuition of a minimum-dimensional space that preserves the distance relationships. One recent work of Houle et al. [6] focuses on the efficient $\mathcal{NN}$ retrieval in subspaces.

3 Definition of Subspace Nearest Neighbor Search

As outlined in Section 2, the $\mathcal{SNNS}$ problem is related to existing approaches, but differs substantially in some aspects, mainly in the dependency of the subspaces quality and the query. In the following we define specifics of the $\mathcal{SNNS}$ problem and introduce an initial model to identify relevant candidate subspaces.

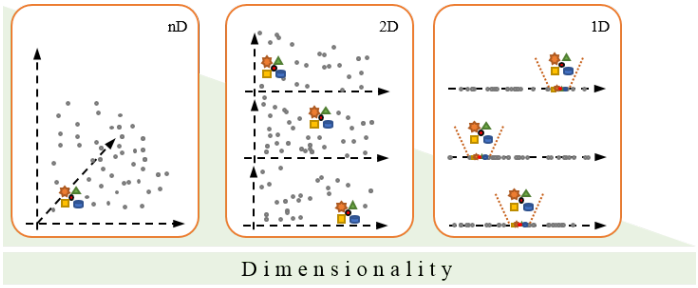


Fig. 2: Illustration of our subspace model: A subspace is considered *relevant*, iff the $\mathcal{NN}$ are also similar to the query in *all* dimensions of the subspace.

The aim of $\mathcal{SNNS}$ can be divided into two coupled tasks: (a) *detect all previously unknown subspaces* that are *relevant* for a $\mathcal{NN}$ search of a *given query*, and (b) determine the respective set of $\mathcal{NN}$ within each relevant subspace. Hereby, the central component is the user-defined query. Different queries may change the relevance of subspaces and affect the resulting $\mathcal{NN}$-sets. Therefore, the characteristics of the query need to be considered for the subspace search strategy (Section 4) and the subspace quality criterion (see below).

We propose an initial subspace model[5] to derive the relevance of a subspace w.r.t. a $\mathcal{NN}$-search. As illustrated in Fig. 2, a subspace is considered *relevant*, iff the following holds: "A set of objects a, b, c are $\mathcal{NN}$ of the query q in a subspace s, iff a, b, and c are a $\mathcal{NN}$ of q in *all* dimensions of s." More formally:

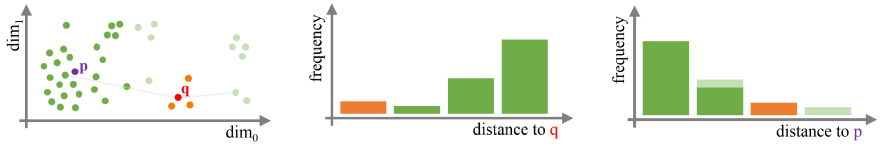
$$\forall_{n \in nn(q,s)} \; and \; \forall_{d \in dim(s)} : \; n \in nn(q,d)$$

whereby $nn(q,s)$ indicates the $\mathcal{NN}$ of q in s, and $dim(s)$ the set of dimensions of the subspace. This principle of a *common set* of $\mathcal{NN}$ in different dimensions is similar to the concept of the *shared nearest neighbor distance* [5] or similar concepts from Data Mining such as consensus clustering or consensus ensembles. The intuition is that the member dimensions of a subspace agree (to a certain minimum threshold) in their $\mathcal{NN}$ rankings when considered individually.

The proposed *item-based* concept is different to the distance distribution-based model presented in [4], or most subspace clustering approaches. Besides the advantage of a semantic $\mathcal{NN}$ interpretability, the model allows to compute heterogeneous subspaces. The relevance of a subspace is independent of a global distance function, but relies on a $\mathcal{NN}$ computation in all dimensions (c.f. Section 4).

Not every subspace, considered relevant by our model, is necessarily *interesting* in all application scenarios. In the medical example from the beginning, a physician will focus on the semantic interpretability of the results, while accepting potential redundancy information. In other scenarios, the minimal description of a subspace may be preferred (c.f. intrinsic dimensionality [3]). Alternative interestingness definitions, such as focusing on subspaces with a minimum –respectively maximum– number of $\mathcal{NN}$ could be possible, too. Generally, the *quality criterion* for nearest neighbor subspaces, has to be regarded application dependent.

[5] Our model assumes *axis-parallel* subspaces. Further research is necessary to analyze the usefulness of *arbitrarily-oriented* subspaces for $\mathcal{NN}$ search.



(a) Data distribution. (b) Characteristic dim. (c) Non-characteristic dim.

Fig. 3: Distance distribution based measure to determine the characteristic of a dimension w.r.t. a $\mathcal{NN}$ search of a given query.

4 Discussion and Open Research Questions

While initial experiments[6] hint on the usefulness of $\mathcal{SNNS}$, we have identified six central research directions that should be explored in the future.

Determine $\mathcal{NN}$ per Dimension. A central question that arises from the model definition is when a data record is considered as $\mathcal{NN}$ to q. Whenever similarity is modeled by a distance function we need to define, detect, or learn an appropriate $\mathcal{NN}$ membership threshold. In our initial experiments, we found that relative threshold definitions should be preferred over fixed thresholds (e.g. a fixed threshold $k = 10$, but the top 20 records are highly similar).

Search Strategy. The number of axis-parallel subspaces is $2^d - 1$ for a dataset with d dimensions. Consequently, an efficient search strategy is necessary to quickly detect interesting subspaces. *Top-down* approaches, based on a so-called *locality criterion* [9], assume that relevant subspaces can be approximated in the full-dimensional space. Yet, our initial experiments lead to the assumption that shared $\mathcal{NN}$ in independent dimensions, as required by our model, can only be found by a *bottom-up* strategy that starts from $\mathcal{NN}$ in individual dimensions. Our model fulfill the *downward closure property* [9] which can boost the search strategy by make use of *APRIORI-like* algorithms.

Query-Based Interestingness for Dimensions. The search strategy can further benefit by focusing on interesting dimensions. We propose an measure for single dimensions, based on the idea described [4], that extracts the characteristic of dimension w.r.t. the query. As shown in Fig. 3, dimensions in which most data records are similar than the query are considered as non-characteristic, hence they are less interesting for possible subspaces.

Subspace Quality Criterion. As indicated in Section 3, novel query dependent quality criteria are necessary to evaluate and rank the interestingness of all detected subspaces. The intuition to measure a subspace's quality differs significantly from earlier approaches, as outlined in Section 2.

Evaluation. Further evaluation schemes for assessing the quality of $\mathcal{SNNS}$ result need to be developed. As described in [12], evaluating subspace analysis methods is difficult, as no real-world dataset with annotated subspace information exists. Likewise, synthetic datasets, such as applied in the evaluation of subspace clustering (e.g. OpenSubspace Framework [12]), cannot be used due to the different analysis goals (c.f. Section 2). Hence, research will benefit from a central, established ground-truth dataset for the evaluation of $\mathcal{SNNS}$.

[6] C.f. supplementary material in the appendix.

Multi-Input $\mathcal{SNNS}$. In many scenarios such as in the medical domain, a small set of query records needs to be investigated by means of $\mathcal{SNNS}$. One challenge for *multi-input* $\mathcal{SNNS}$ are dimensions in which the set of queries differ.

5 Conclusion

This position paper outlines a novel research problem, called subspace nearest neighbor search, which aims at determining *query-dependent* subspaces for nearest neighbor search. We delineated subspace nearest neighbor search from existing approaches, described an initial model to retrieve relevant subspaces, and identified central research directions to explore in the future. Initial experiments have proven the usefulness and that it is beneficial to drive research in this field.

Acknowledgments. The research leading to these results has received funding from the "SteerSCiVA: Steerable Subspace Clustering for Visual Analytics" DFG-664/11 project.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE TKDE 17(6), 734–749 (2005)
2. Beyer, K.S., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is "nearest neighbor" meaningful? In: Proc. 7th Int. Conf. Database Theory. pp. 217–235 (1999)
3. Camastra, F.: Data dimensionality estimation methods: a survey. Pattern Recognition 36(12), 2945–2954 (2003)
4. Hinneburg, A., Keim, D.A., Aggarwal, C.C.: What is the nearest neighbor in high dimensional spaces? In: Proc. 26th Int. Conf. on VLDB, Cairo, Egypt (2000)
5. Houle, M.E., Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Can shared-neighbor distances defeat the curse of dimensionality? In: Scientific and Statistical Database Management. pp. 482–500. Springer (2010)
6. Houle, M.E., Ma, X., Oria, V., Sun, J.: Efficient algorithms for similarity search in axis-aligned subspaces. In: SISAP. pp. 1–12. No. 8821 (2014)
7. Johansson, S., Johansson, J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. IEEE TVCG 15, 993–1000 (2009)
8. Kailing, K., Kriegel, H.P., Kröger, P., Wanka, S.: Ranking interesting subspaces for clustering high dimensional data. In: 7th Proc. of Knowledge Discovery in Databases: PKDD. pp. 241–252 (2003)
9. Kriegel, H.P., Kröger, P., Zimek, A.: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM TKDD 3(1), 1 (2009)
10. Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman & Hall/CRC Press Data Mining and Knowledge Discovery Series (2007)
11. Micenkova, B., Dang, X.H., Assent, I., Ng, R.: Explaining outliers by subspace separability. In: 13th. IEEE ICDM. pp. 518–527 (2013)
12. Müller, E., Günnemann, S., Assent, I., Seidl, T.: Evaluating clustering in subspace projections of high dimensional data. In: Proc. of 35th Int. Conf. on Very Large Data Bases. vol. 2, pp. 1270–1281 (2009)
13. Zimek, A., Schubert, E., Kriegel, H.P.: A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining 5(5), 363–387 (2012)

Supplementary Material

In this section, we provide supplementary material, that we will publish on a website, in case this paper is accepted at the SISAP conference.

Initial Experiments

We performed two initial experiments to show the advantages of subspace nearest neighbor search. In the first experiment, we compare the $\mathcal{NN}$ of a Euclidean distance-based full-dimensional query with $\mathcal{NN}$ in different subspaces, computed by our subspace model. In the second experiment, we compare the different characteristics of a dimension for different queries. In the following, we describe the underlying dataset and give details on the two experiments.

Dataset: Nutrition Database We applied our experiments on a real-world dataset called USDA National Nutrient Database for Standard Reference[7]. This database is the main source of food composition data in the United States. It contains different food items such as different types of *milk products*, *vegetables*, *fruits* and *meat*, but also complete food products such as *baby food containing meat* or *Kellog's cereals*. Each food item is described by about 50 different numerical attributes such as the amount of *water*, *energy*, *vitamin c*, *calcium*, or *sodium*.

For the following two experiments, we selected the food items *butter with salt* and *gauda cheese* as queries.

Experiment 1: Comparison of $\mathcal{NN}$ in Different Subspaces
In our first experiment, we compared the $\mathcal{NN}$-sets of different subspaces. In order to do so, we first computed the set of $\mathcal{NN}$ based on the Euclidean distance in the full-dimensional space and afterwards applied our model to extract several relevant subspaces together with its $\mathcal{NN}$-sets. The tables 1 and 2 show the different subspaces and nearest neighbors for the queries *butter with salt* and *gauda cheese*.

Butter With Salt. When analyzing the results we can see that in the full-dimensional space, some of the $\mathcal{NN}$ seem not to be similar to butter, i.e. *pancakes*, *kellog's fruit bars*, *tabasco sauce*, and *peppers*. On the other hand, the extracted subspaces reveal some interesting knowledge: $Subspace_1$ contains different kinds of oil (*soya*, *coconut*, *olive*, *sunflower*, etc.) and fat (such as *lard* and *margarine*). The $\mathcal{NN}$ of $subspace_2$ differs a lot. Many items are sweet products such as *candies*, *corn syrup*, and *pudding*.

[7] http://ndb.nal.usda.gov

Gauda Cheese. Both, in the full-dimensional space and in $Subspace_4$, different cheeses are retrieved as nearest neighbors. $Subspace_3$ is interesting, as it contains only two different kind of cheeses, namely *gauda* and *edamer* cheese along with different beef parts. This subspace shows that different types of *beef steaks*, *eyes*, and *loins* are pretty similar to *gauda cheese* in the respective dimensions.

Both examples show that in some applications (e.g. butter example) a subspaces search is necessary in order to retrieve useful $\mathcal{NN}$. Especially the $\mathcal{NN}$ in the full-dimensional space seems not similar to the query which is probably caused by the large number of dimensions that influence the distance computation (c.f. curse of dimensionality [2]). Furthermore, the examples show that subspace nearest neighbor search can help to find objects that are similar to the query in only a subset of dimensions.

Experiment 2: Query-Based Interestingness for Dimensions
In our second experiment, we compared the characteristics of different dimensions to a given query. For each dimension, we computed the distance between each food item and the query, sorted the distances, and created a distance-frequency histogram for each dimension. The underlying idea is similar to the interestingness measure for subspaces described in [4]. Fig. 3 illustrates the idea of the histograms in more detail.

The histograms for all dimensions and for the queries *butter with salt* and *gauda cheese*, can be found in Figures 4 and 5. In these figures, we can clearly see that for a given query the characteristics of dimensions differ significantly. Also, for different queries, different dimensions are characteristic. As described above, the characteristic or interestingness of a single dimension can be considered in the subspace search strategy by e.g. filtering dimensions in order to reduce the search space. Our initial experiments show that the characteristic of dimensions differs significantly. Further research needs to be done in order to find an appropriate filtering technique for dimensions.

Table 1: This table shows the nearest neighbors of 'butter with salt' based on the Euclidean distance in the full-dimensional space, and the nearest neighbors according to our model in two different subspaces: $Subspace_1$ (*Copper*, *GmWt_-1*, *Lipid_Tot*, *Magnesium*, *Phosphorus*, *Potassium*, *Riboflavin*, *Thiamin*), and $Subspace_2$ (*Copper*, *Magnesium*, *Phosphorus*, *Potassium*, *Riboflavin*, *Water*).

Full Space	**Subspace 1**	**Subspace 2**
butter, with salt	butter, with salt	butter, with salt
butter, whipped	butter, whipped	butter, whipped
butter, without salt	butter oil, anhydrous	butter, without salt
butter oil, anhydrous	butter, without salt	salad drsng, mayo
kellogg's, fruit bars	lard	margarine
margarine	salad drsng, mayo	chicken, broilers
pancakes	oil, soybn	pork, backfat
waffle	oil, cocnt	candies, butterscotch
cream	oil, olive	candies, hard
cheese, cream	oil, safflower	candies, jellybeans
pie crust	vegetable oil, palm kernel	candies, mars snackfood
cheese, mozzarella	oil, canola	chewing gum
kellogg's cereals	oil, sunflower	puddings, vanilla
soup	margarine	jellies
cheese, limburger	shortening	sweeteners, tabletop
peppers	chicken, broilers	syrups, corn
sauce tabasco	oil, corn, peanut, and olive	syrups, maple

Table 2: This table shows the nearest neighbors of 'gauda cheese' based on the Euclidean distance in the full-dimensional space, and the nearest neighbors according to our model in two different subspaces: $Subspace_3$ (*Copper*, *Panto_-Acid*, *Protein*, *Vit_B12*, *Vit_E*, *Zinc*), and $Subspace_4$ (*Cholestrl*, *Copper*, *Energ*, *FA_Mono*, *FA_Sat*, *Lipid_Tot*).

Full Space	**Subspace 3**	**Subspace 4**
cheese, gouda	cheese, edam	cheese, brick
cheese, edam	cheese, gouda	cheese, brie
cheese, provolone	beef, t-bone steak	cheese, cheddar
cheese, monterey	beef, steak, bnless	cheese, edam
cheese, provolone	beef, top loin	cheese, gouda
cheese, mozzarella	beef, steak	cheese, limburger
cheese, feta	beef, eye of rnd	cheese, monterey
cheese, mexican	beef, prtrhs steak	cheese, mozzarella
cheese, fontina	–	cheese, muenster
cheese, camembert	–	cheese, parmesan
cheese, mexican	–	cheese, port de salut
cheese, blue	–	cheese, romano
cheese, cheddar	–	cheese, swiss
cheese, colby	–	cheese, american fort

Fig. 4: The distance-frequency histograms for all dimensions w.r.t. the query *butter with salt*.

Fig. 5: The distance-frequency histograms for all dimensions w.r.t. the query *gauda cheese*.