Subspace Nearest Neighbor Search - Problem Statement, Approaches, and Discussion

Position Paper

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

Abstract. Computing the similarity between objects is a central task for many applications in the field of information retrieval and data mining. For finding k-nearest neighbors, typically a ranking is computed based on a predetermined set of data dimensions and a distance function, constant over all possible queries. However, many high-dimensional feature spaces contain a large number of dimensions, 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 or -weighting typically provide a global subspace selection, which may not be suitable for all possibly queries. In this position paper, we frame a new research problem, called *subspace nearest neighbor search*, aiming at multiple query-dependent subspaces for nearest neighbor search. We describe relevant problem characteristics, relate to existing approaches, and outline potential research directions.

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 a distance function. However, the consideration of all dimensions and a single distance function may not be appropriate for all queries, as we will discuss in the following.

For datasets with a high number of dimensions, similarity measures may loose their discriminative ability since 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 instability increases with the proportion of irrelevant or conflicting dimensions.

Consider the following clinical example: 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 is confronted with a high number of unrelated diseases and respective symptoms. The most similar patients (nearest neighbors, $\mathcal{NN}$) based on all features are often not suited to guide the diagnostic process as irrelevant dimensions, such as the hair color, may dominate the search process. Meaningful conclusions can only be drawn if the *characteristic* dimensions for the particular 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 exist? Many other application examples can be found, where $\mathcal{NN}$ search in query-dependent subspaces is potentially relevant, e.g., 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 above example, 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.

$\mathcal{SNNS}$ is inspired by works in subspace clustering and -search. However, it differs from these fields, as the goal is to derive query-dependent subspaces. Therefore, we define a novel problem definition. In $\mathcal{SNNS}$, our goal is to (1) detect *query dependent* and *previously unknown subspaces* that are relevant, and (2) derive the corresponding nearest neighbor set to the query within that corresponding subspace. This paper addresses the following questions: "What is a relevant subspace for a given query?", "How can we computationally extract this relevance information?", and "How can we adapt ideas from subspace clustering, outlier detection, or feature selection for $\mathcal{SNNS}$?"

2 Related Problems

Next, we give a concise overview of the fields related to $\mathcal{SNNS}$. An overview about the fields and its relation to $\mathcal{SNNS}$ 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 (e.g., classification error). As shown in Ⓑ, there are two main differences to $\mathcal{SNNS}$: Feature selection derives a single subspace (result view) for all analysis tasks, and the resulting subspace is query independent. In contrast, $\mathcal{SNNS}$ is aiming at a *faceted result view* of multiple, query-dependent subspaces.

Subspace Clustering. Subspace clustering aims at finding clusters in different axis-parallel or arbitrarily-oriented subspaces [9]. The approaches are based on

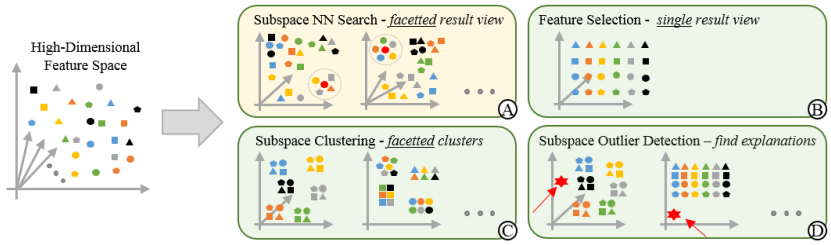

Fig. 1: Focus of Subspace Nearest Neighbor Search ($\mathcal{SNNS}$) and related approaches: While $\mathcal{SNNS}$ aims at multiple, query-dependent subspaces, related fields focus on a single result or on subspaces with different properties.

subspace search methods and heuristics to measure the subspace cluster quality. The computation of clusters and subspaces can be tightly coupled or decoupled, see e.g., [8]. As shown in Ⓒ, subspace clustering and $\mathcal{SNNS}$ both aim at a facetted result, but differ in their relevance definition of a subspace: dense clusters vs. query-dependent nearest neighbors in multiple subspaces.

Subspace Outlier Detection. Methods in this area search for subspaces in which an arbitrary, or a user-defined object is considered as outlier [13]. As before, the search process consists of subspace search methods and criteria to measure the subspace quality, 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 definition of a subspace differs significantly as $\mathcal{SNNS}$ searches for objects that are similar to the query, while subspace outlier detection seeks for objects dissimilar to all other objects.

Query-dependent Subspace Search. In [5] it was proposed to determine one query-dependent subspace to improve $\mathcal{NN}$-queries. The authors describe an approach to measure the quality of a subspace by the separability between all data records and the $\mathcal{NN}$ of a query. In their evaluation, they show that a query-dependent subspace reduces the error of a $\mathcal{NN}$-classification substantially. The work can be seen as initial approach on $\mathcal{SNNS}$ and, therefore, most closely relates to our work. However, the general aims of [5] differ, as it does not search for a facetted result view, i.e. different $\mathcal{NN}$ sets 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 other recent work focuses on the efficient $\mathcal{NN}$ retrieval in subspaces [7].

3 Definition of Subspace Nearest Neighbor Search

In the following we define characteristics of the $\mathcal{SNNS}$ problem and introduce an initial model to identify relevant candidate subspaces.

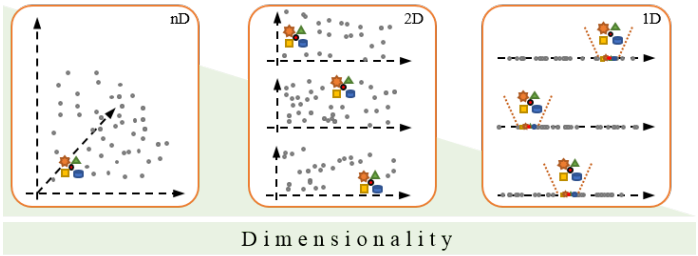

Fig. 2: Illustration of our subspace model: A subspace is considered *relevant*, iff the nearest neighbors are similar to the query in *all* dimensions of the subspace.

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. 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 and the evaluation criterion (c.f. Section 4).

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* [6] or consensus methods. The intuition is that the member dimensions of a subspace agree (to a certain minimum threshold) in their $\mathcal{NN}$ rankings, when considered individually.

This *item-based* subspace concept is different to the distance distribution-based model presented in [5], 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 individual $\mathcal{NN}$ computations in all dimensions.

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 redundant 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 as 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 queries p and q.

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.

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

Query-Based Interestingness for Dimensions. The subspace search strategy can further benefit by focusing on interesting dimensions. We propose a measure for single dimensions, based on the idea described in [5] that extracts the characteristic of dimension w.r.t. the query. As shown in Fig. 3, dimensions in which most data records are similar to the query are considered as non-characteristic, hence they are less interesting for possible subspaces.

Subspace Quality Criterion. Novel criteria are needed to rank the detected subspaces by their interestingness. The intuition to measure a subspace's quality differs significantly from earlier approaches, as outlined in Section 2. In addition, novel user interfaces and visualizations are necessary to understand and interpret multiple, partially redundant, subspaces and their different rankings [4].

Evaluation. Evaluating subspace analysis methods is challenging, as obtaining real-world dataset with annotated subspace information is expensive [12]. Likewise, synthetic data for the evaluation of subspace clustering (e.g., *OpenSubspace Framework* [12]), differs in the analysis goals (c.f. Section 2). Hence, research will benefit from a established ground-truth dataset for the evaluation of $\mathcal{SNNS}$.

[6] C.f. supplementary material on our website: `http://files.dbvis.de/sisap2015/`.

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 ($\mathcal{SNNS}$), which aims at determining *query-dependent* subspaces for nearest neighbor search. Initial experiments have proven the usefulness and that it is beneficial to drive research in this field.

Acknowledgments. We would like to thank the German Research Foundation (DFG) for financial support within the projects A03 of SFB/Transregio 161 "Quantitative Methods for Visual Computing" and DFG-664/11 "SteerSCiVA: Steerable Subspace Clustering for Visual Analytics".

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. Gleicher, M., Albers, D., Walker, R., Jusufi, I., Hansen, C.D., Roberts, J.C.: Visual comparison for information visualization. Information Visualization 10(4), 289–309 (2011)
5. 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)
6. 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)
7. 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)
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: VLDB. 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)