Kaya, Muhammed-Fatih; Schoop, Mareike

Analytical Comparison of Clustering Techniques for the Recognition of Communication Patterns

Group Decision and Negotiation

Provided in Cooperation with:
Springer Nature

Suggested Citation: Kaya, Muhammed-Fatih; Schoop, Mareike (2021) : Analytical Comparison of Clustering Techniques for the Recognition of Communication Patterns, Group Decision and Negotiation, ISSN 1572-9907, Springer Netherlands, Dordrecht, Vol. 31, Iss. 3, pp. 555-589, https://doi.org/10.1007/s10726-021-09758-7

Group Decision and Negotiation (2022) 31:555–589
https://doi.org/10.1007/s10726-021-09758-7

Analytical Comparison of Clustering Techniques for the Recognition of Communication Patterns

Muhammed-Fatih Kaya[1] · **Mareike Schoop**[1]

Accepted: 22 August 2021 / Published online: 13 October 2021

Abstract

The systematic processing of unstructured communication data as well as the milestone of pattern recognition in order to determine communication groups in negotiations bears many challenges in Machine Learning. In particular, the so-called curse of dimensionality makes the pattern recognition process demanding and requires further research in the negotiation environment. In this paper, various selected renowned clustering approaches are evaluated with regard to their pattern recognition potential based on high-dimensional negotiation communication data. A research approach is presented to evaluate the application potential of selected methods via a holistic framework including three main evaluation milestones: the determination of optimal number of clusters, the main clustering application, and the performance evaluation. Hence, quantified Term Document Matrices are initially pre-processed and afterwards used as underlying databases to investigate the pattern recognition potential of clustering techniques by considering the information regarding the optimal number of clusters and by measuring the respective internal as well as external performances. The overall research results show that certain cluster separations are recommended by internal and external performance measures by means of a holistic evaluation approach, whereas three of the clustering separations are eliminated based on the evaluation results.

Keywords Communication data · Pattern recognition · Clustering · Optimisation · High-dimensional data · Term document matrix

✉ Muhammed-Fatih Kaya
Muhammed-Fatih.Kaya@uni-hohenheim.de

Mareike Schoop
Schoop@uni-hohenheim.de

[1] University of Hohenheim, 70599 Stuttgart, Germany

1 Motivation

Negotiations and communication are inherently intertwined. Negotiators communicate to point out their individual positions, to exchange information and strategies, and to enter into business relationships (Weingart and Olekalns 2004). Hence, negotiation cannot occur without some means of communication (Putnam and Roloff 1992); communication reflects the behaviour of the negotiating participants that can take different forms (Donohue and Roberto 1996; Hargie and Dickson 2004). If negotiations are conducted electronically, communication is particularly important as other non-verbal cues such as mimics, gestures, tone of voice are not there to help interpret the message (Croson 1999; Purdy et al. 2000; Schoop 2021). This leads to the need for careful analysis of electronic communication interactions to ensure a common understanding between the negotiating parties.

The determination of patterns in communication data serves exactly this purpose of categorising exchanged communication interactions into systematic pattern groups to enable further in-depth analysis of communicative interactions. Patterns may be hidden in communications but can reveal important information about the communicative behaviour (Donohue and Roberto 1996; Sokolova et al. 2004). However, the implementation of pattern recognition poses numerous challenges mainly due to the unstructured underlying data. The unstructured nature of data triggers the so-called "curse of dimensionality" which is caused by transforming the unstructured data into structured processable data. Complex pre-processing and transformation steps are called for to overcome these challenges which prevents an easy automated analysis of negotiation communication (Yan et al. 2006). The analysis of high-dimensional data additionally requires an extensive pre-evaluation of techniques, as common statistical methods are not suitable for this kind of complex cases (Kumar 2009). Consequently, various analysis methods from Machine Learning have to be considered to extract value-adding information from unstructured communication data in addition to substantial preparation effort.

Communication particles e.g. in the form of negotiation messages or sentences have so far been primarily examined by a manual coding-based approach to avoid the processing effort. Nevertheless, this approach has some significant disadvantages. The feasibility can be problematic due to the high effort since the human coders have to split the given communication into single codable units and subsequently assign them to communication-based category groups derived from theory which is both complex and subjective (Weingart et al. 2004). Moreover, only a certain amount of data (which is defined in advance and is convertible for human coders) can be processed due to the manual processing of communication units. The potential for success of machine-based techniques for processing large amounts of data must consequently be taken into account with great care in an age of rapid data floods. Data types of an unstructured and textual nature represent nearly 80% of the data volume (Das and Kumar 2013).

The current research paper will, therefore, use Machine Learning methods to evaluate whether and to what extent renowned clustering techniques are suitable

for the recognition of groups of patterns by considering high-dimensional communication data. The research question is as follows:

> How do renowned clustering methods perform with regard to detection of pattern groups in high-dimensional negotiation communication data?

To answer the research question, chapter 2 provides the theoretical background by emphasising the special importance of communication in the application context of negotiations. Furthermore, the challenges of clustering techniques in high-dimensional space are presented related to the processing of textually exchanged messages and the so-called curse of dimensionality. In chapter 3, the generated research approach is presented which calls for an iterative evaluation of selected clustering techniques necessary to answer the research question. The evaluation results are introduced in chapter 4 and are subsequently subjected to an analytical comparison in the following discussion chapter (chapter 5). The research paper concludes with a summary of the key findings and a research outlook.

2 Theoretical Background

2.1 The Importance of Communicative Interactions in Electronic Negotiations

In view of the increasing digitalisation, the importance of electronic negotiations has increased significantly making it a topic of high strategic relevance for many companies (Lewicki et al. 2016), e.g. due to the possibility of global trade interactions with dislocated and asynchronous digital process support.

Negotiations are defined as an iterative process of communication and decision-making between at least two parties that are unable to achieve their personal negotiating objectives through unilateral actions (Bichler et al. 2003). Negotiating parties try to reach a compromise by striving for a common consensus. In doing so, they deal with negotiation issues that may be intertwined and they exchange key information during the negotiation process through the use of communication. Arguments, requests, offers, and counter-offers are exchanged in order to realise the different negotiation objectives. Finally, a negotiation is terminated with a final acceptance or rejection depending on the course of the negotiation and the reactions of the negotiating partners (Adair and Brett 2005).

There is no negotiation without communication (Schoop 2020). Offer and non-offer communication are part of negotiation processes (Tutzauer 1992); decision-making can only be ensured by using an efficient communication process (Putnam and Roloff 1992). Communication enables the exchange of information, preferences, and strategies as well as a sequence of offers and establishes long-term relationships between the negotiating parties. Furthermore, communication is also used to manage the negotiation process itself (Putnam 2010; Weingart and Olekalns 2004). From a more general point of view, communication represents a central link to ensure the coordination between the partners and implicitly reflects the individual negotiation behaviour of each party (Myers and Myers 1982; Rogers and Agarwala-Rogers

1976). The transmitted communication behaviour represents negotiation tactics (Donohue and Roberto 1996).

The communication channel is of particular importance due to the absence of non-verbal communication especially in e-negotiations since the exchange of communication and in particular the negotiation messages are exclusively carried out via digital channels (Hargie and Dickson 2004). The absence of so-called social cues (e.g. facial expressions, gestures, mimics or tone of voice) can make the interpretation of utterances more difficult which in turn can lead to increased misunderstandings between the negotiating parties. Electronic communication thus requires specific support in the form of a content-based enrichment to account for missing cues and to provide different cues.

Negotiation Support Systems (NSSs) assist complex negotiations interactions by providing essential support to the decision-making and communication process (Schoop et al. 2003; Schoop 2020). Communication support is of particular importance since the exchange of information as well as individual preferences are performed by means of communication with the above challenges of using digital media. It is, therefore, of utmost importance that negotiating parties aim for joint understanding and that the NSS supports that goal (Schoop 2021; Vetschera et al. 2011).

The NSS Negoisst is the only NSS that has a communication-centred approach. It is based on Speech Act Theory by Searle et al. (1969) and on the Theory of Communicative Action by Habermas (1981) which define the factors of understanding and of good communication. Negoisst supports communication and understanding on all three semiotic levels, namely syntactics, semantics, and pragmatics (Morris 1971; Schoop 2010, 2021).

Negotiation communication in Negoisst is based on a negotiation protocol that manages the negotiation process. Communication is conducted via asynchronous messages using natural language (Schoop 2004). A negotiation always starts with an opening initial offer immediately followed by a series of counteroffers generated from the negotiating partners. This continues until either a common consensus can be reached which can be the final acceptance and thus a contract deal or the final rejection and thus the unsuccessful end of the process. Negoisst also offers informal communication by means of questions and answers in order to clarify ambiguities or to discuss ideas without commitment (Schoop 2010). The message exchange is enriched by providing a message type representing the communication mode (namely offer, counteroffer, question, clarification, acceptance, rejection) and by linking the unstructured text to a structured negotiation vocabulary and negotiation agenda. The former is the pragmatic enrichment whereas the latter is the semantic enrichment (Schoop et al. 2014).

Even though Negoisst offers such extensive communication support, further aspects could enhance the negotiation support even more. In particular the restructuring and subsequent machine-based categorisation of unstructured communication data using pattern recognition methods of Machine Learning could transfer high-dimensional negotiation messages into machine-determined groups of communication behaviour. Hence, this kind of support functionality would allow the negotiating units to be systematically separated pattern by pattern, subsequently categorised

and finally made available to the data science analyst for further interpretation and processing. Nonetheless, the automatic processing of unstructured negotiation communication faces a number of challenges which must be overcome especially in the negotiation environment. This is the only way to recognise useful patterns in communication that contribute to the support character of NSSs.

2.2 Clustering of High-Dimensional Negotiation Messages

Negotiation messages in the NSS Negoisst are text messages that are exchanged bilaterally. Therefore, they are unstructured in nature and enriched by various means as described in the previous section. They provide important indicators of strategic orientation and emotional state and represent the overall negotiation behaviour (Hargie 2010). In contrast to structured data, unstructured data—in our case communication data—must be subjected to extensive pre-processing steps in order to derive value-added knowledge for structured categorisation (Feldman and Sanger 2007; Vijayarani et al. 2015). The overall goal of the processing is to find a structured numerical representation of the underlying communication data in the form of a vector space model. Such model maintains the richness of the data and can be further processed in additional data science steps to derive new knowledge (Kaya and Schoop 2020).

The curse of dimensionality poses a major challenge for the structuring steps, especially at the milestone of transformation, and must, therefore, be subjected to iterative evaluations as otherwise important indicators may be lost in the process of dimensionality reduction (Kadhim et al. 2014). These indicators are of particular importance to our defined research goal and represent a significant influencing factor, namely the application of clustering techniques for the determination of categorical patterns in high-dimensional communication data. In addition, appropriate methods for categorising the underlying communication data must be considered which are able to operate on a high-dimensional data extent.

Clustering describes a Machine Learning approach in which objects are systematically grouped into a number of categorical groups based on similarity; these groups are called clusters (Gan et al. 2007). Objects assigned to the same cluster should be as similar as possible and objects assigned to different clusters should be as dissimilar as possible (Huang 2008). The ability to identify previously unknown groups and patterns in existing data through clustering makes it a very useful tool in a variety of application fields and for many research directions (Frades and Matthiesen 2010).

Clustering techniques can be divided into partitioning, hierarchical clustering and density-based clustering (see Fig. 1). *Partitioning* divides the objects of the dataset or vectors into a certain number of partitions whereby one partition corresponds to a cluster (Reynolds et al. 2006). The number of clusters is usually pre-defined. Furthermore, an objective function is to be optimised by forming the clusters and minimising the distance between objects in the same cluster (Saket and Pandya 2016). Hence, the number of clusters must firstly be determined and evaluated by the user in a heuristic way followed by an iterative procedure for optimisation (Dharmarajan and Velmurugan 2013; Shah and Mahajan 2012). Partitioning clustering

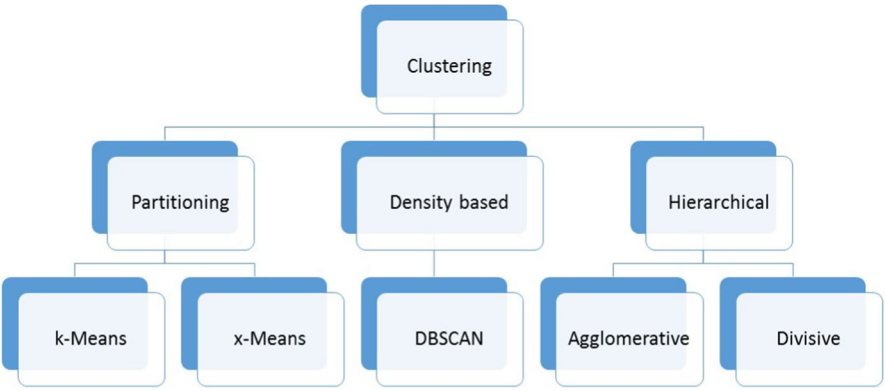

Fig. 1 Overview of clustering techniques (cf. Santhisree and Damodaram 2011; Saxena et al. 2017)

techniques have been considered in several recent research works such as visualisation, health care, document structuring (Abualigah et al. 2016; Allahyari et al. 2017; Silitonga 2017). *Hierarchical* clustering methods divide the objects of a dataset into a sequence of nested partitions which results in a hierarchy (Gan et al. 2007). The result of this approach can be visualised as a tree called a dendrogram that presents the relations between the clustering groups. This is an essential difference to partitioning methods that divide objects from the same categorisation level into individual partitions to form clusters. In addition, the relational dependency between the determined clusters cannot be traced back with the partitioning approach (Davidson and Ravi 2005). Hierarchical procedures play a crucial role in the effective application of clustering especially to textual datasets and document clustering (Bafna et al. 2016; Renganathan 2017; Shehata et al. 2006). *Density-based* clustering methods enable the determination of clusters of arbitrary shape (Ester et al. 1996; Zhu et al. 2016). To do so, such approaches make assumptions about either the density or the variance of the underlying data and take the radius of the neighbourhood for each clustering object and the minimum number of objects that may form a cluster to determine cluster sizes (Kriegel and Pfeifle 2005; Kriegel et al. 2011).

The application potential of the described clustering approaches is manifold. Hierarchical as well as partitioning clustering approaches are used most often. Whilst hierarchical approaches often achieve better results, partitioning methods compute faster (Fred and Leitao 2000). Especially in high-dimensional space, experimental results have shown that partitioning clustering approaches are well suited for large document datasets and generate stable clustering performances due to their relatively low computational requirements (Shah and Mahajan 2012). Nevertheless, they pose the problem of pre-defining the number of clusters which might be challenging especially in high-dimensional space (Kodinariya and Makwana 2013; Rokach and Maimon 2005). On the other hand, density-based methods are among the most renowned clustering methods and are popular due to their high application potential for high-dimensional data (Chen et al. 2018; Tran et al. 2006). They are

less sensitive to outliers and are also able to determine clusters with different shapes (Bhagat et al. 2016).

In summary, the various clustering approaches have different advantages and disadvantages and must therefore be carefully evaluated with regard to the application potential and success potential especially for high-dimensional data. Negotiation communication datasets must be divided into cluster groups with maximum accuracy by taking into account their characteristic properties, since they might contain important indicators of negotiation behaviour. In addition, the field of semi-automated clustering based on communication data is unexplored so far and offers high potential in terms of pattern recognition.

3 Research Approach

Due to the unstructured nature of communication data and the associated curse of dimensionality, clustering requires some pre-processing steps. To carry out a holistic consideration for the application of cluster variants, a research approach is now presented which reflects the procedural as well as the methodological implementation of this research paper milestone by milestone (see Fig. 2).

For a systematic structuring of unstructured data using clustering methods, a prepared database must be available to take the maximum information content into account for carrying out efficient clustering procedures (Mirkin 2012; Zerhari et al. 2015). Thus, some pre-processing steps are necessary before the active implementation of clustering can be conducted, so that the first milestones are

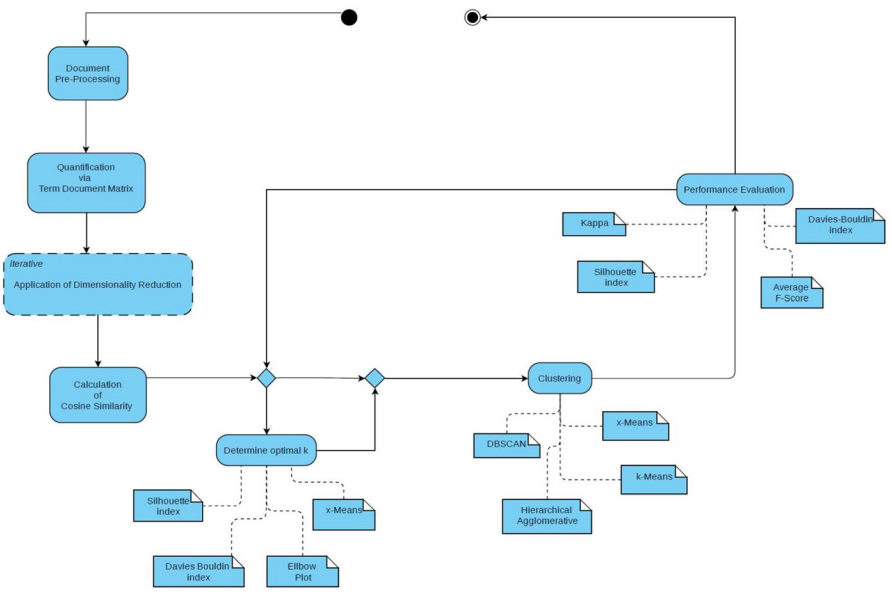

Fig. 2 Research approach (cf. Kadhim et al. 2014; Rana et al. 2019)

represented by the achievement of structured datasets. Firstly, unstructured communication data is transformed into structured term document matrices and data is quantified by means of composite vector space models based on the TF-IDF and frequency measures. Whilst the frequency measure calculates the relative frequency of each word in the document (here: negotiation sentence), TF-IDF mathematically represents the product of the term frequency, the inverse document frequency and weights how important the particular word in the investigated document is (Munot and Govilkar 2014). The data is subjected to a series of preprocessing steps from classical text mining in subsequent steps (Erk 2012; Munková et al. 2013).

3.1 Dimensionality Reduction

As described in the previous section, the quantification of the underlying natural language leads to the curse of dimensionality which represents a major challenge for the pre-processing steps as well as for the clustering techniques to be performed. Therefore, this dimensional expansion has to be evaluated extensively using different reduction algorithms in an iterative manner until a database can be found that reproduces the underlying information and indicators with a reduced set of dimensions (Kaya and Schoop 2020). For dimension reduction, three Feature Selection (FS) methods, two Feature Extraction (FE) methods and a statistical approach are used. Whilst the latter determines correlating term dimensions statistically and eliminating them accordingly, FS and FE use machine learning methods to perform intelligent dimensionality reduction (Khalid et al. 2014; Zebari et al. 2020). In FS, those dimensions are selected from a set of dimensions that significantly contribute to the model performance (Li et al. 2017). Forward Selection is an FS approach; it starts with an empty set and gradually adds relevant dimensions to the set of selected dimensions until an increase in performance can no longer be achieved (Gheyas and Smith 2010). Backward Elimination starts with the entire set of attributes and successively removes those attributes that cause the smallest decrease in terms of performance until there is no more increase in performance (Maldonado et al. 2014). While both approaches face the danger of getting stuck at the local optimum, the third approach called Optimize Selection (OS) uses a greedy algorithm to approximate the local optima to the global optimum in a heuristic way (Venkatesh and Anuradha 2019). In contrast, FE does not aim to find an optimised selection of attributes, but extracts novel dimensions by systematically combining existing attributes so that multiple dimensions are represented by novel composite dimensions (Shah and Patel 2016). The Singular Value Decomposition method aims at reducing those dimensions that show a linear dependence. Principal Component Analysis (PCA) reduces dimensions to a set of factors by means of linear combinations and orthogonal transformations to convert previously correlated attributes to non-correlated attributes (Qu et al. 2002; Wall et al. 2003). The milestone of dimensionality reduction is of fundamental importance and requires an increased effort to be able efficiently implement subsequent procedural milestones.

3.2 Calculation of Similarity Measure

Once the basic data framework is given, a similarity measure must be calculated as part of the subsequent milestone by considering the vector representations of the documents (here: negotiation sentences) based on the TDM. A similarity measure is used for the systematic and accurate grouping of data objects (Irani et al. 2016). There are different similarity criteria for clustering textual communication data. Based on the chosen quantification variant and on literature recommendations, cosine similarity is proposed for our application context since it can be used for efficient clustering of numerically transformed high-dimensional data (Muflikhah and Baharudin 2009; Ravindran and Thanamani 2015). More precisely, the cosine similarity calculates the normalised cosine of an angle between two sophisticated vectors and thus measures the degree of similarity between data objects. From a mathematical perspective, cosine similarity calculates the dot product of two document vectors divided by the product of the vector lengths (Huang 2008). The spectrum of the cosine similarity lies between 0 and 1. The more distant the angles of the document vectors are from each other, the smaller the cosine value. A comparison can be made between the document vectors by comparing these results (Gunawan et al. 2018).

3.3 Evaluation of Optimal Cluster Number

Before the clustering procedure can be applied, the evaluation of the optimal cluster number k is required (see Fig. 2). This step is necessary for all those clustering methods that require a predefined number of clusters as input. Furthermore, such number of clusters also provides important additional insights for approaches that do not require a predefined k to get an idea of the recommended number of clusters. Consequently, this milestone can be interpreted as a first pre-evaluation depending on the approach to be applied. We will now compare the following approaches to determine the number k as accurately as possible for the negotiation communication data at hand: the elbow plot considering the average centroid distance, the sequential evaluation of the Davies Bouldin index, the Silhouette index, and the x-Means approach.

The heuristic elbow method selects the smallest value of k in the generated scree plot at which the distortion starts to increase most sharply so that a strong transition exists at the transition to the previous k-value (i.e. k−1) (Syakur et al. 2018). It chooses the number of clusters in a way that adding another cluster would not result in better data modelling for the underlying dataset. When generating the elbow plot, the average within centroid distance is taken as the criterion to be considered. The compactness of the clusters is numerically represented in this way (Bholowalia and Kumar 2014).

As a further complement, the Davies Bouldin (DB) index is used to assess the optimal number of k. This index is evaluated in an iterative manner based on a predefined interval of k. A corresponding value for the DB is calculated for each cluster number k. More precisely, the DB index is based on the relationship between the distance within each cluster and between clusters (Davies and Bouldin 1979). Thus,

the optimal number of clusters is represented by the lowest DB index minimising the distance within the clusters and maximising the distance between the clusters (Ray and Turi 1999).

The calculation of the Silhouette index represents an approach that is more complex in terms of running time. Compared to the DB index and the elbow method, the Silhouette index provides more information about the quality of the number k and should, therefore, be used as an additional complement (Petrovic 2006). Those clusters that have a positive coefficient and are thus closer to 1 imply that their data points are distant from the neighbouring cluster. These cluster splits are, therefore, to be preferred. Clusters with a coefficient of 0 are very close to the threshold range of the decision boundary between two clusters. On the other hand, negative coefficients indicate a wrong allocation of data points and thus outliers in the corresponding cluster (Aranganayagi and Thangavel 2007). This should be avoided across all clusters, even if the determination of cluster groups cannot always be done unambiguously and precisely in the high-dimensional space. Hence, the Silhouette coefficient is calculated for each k in the given interval similar to the DB index by evaluating the clustering performance using the difference between the clusters and within the clusters via the pairwise distance.

Whilst the DB index selects the minimum value of the evaluation chain, the Silhouette index takes the maximum Silhouette value from the series of results into account (Liu et al. 2010; Rousseeuw 1987).

In addition to the three evaluation methods described, one further approach called x-Means is to be used for the efficient estimation of the number of clusters. It is an extension of k-Means clustering with algorithmic improvements regarding the predictability and the determination of the optimal number. The algorithm uses the BIC index as a splitting criterion by iteratively optimising statistically-based criteria to maximise the model probability in the search for the space of cluster locations (Pelleg and Moore 2000). This approach provides important complementary information for the determination of the optimal cluster number. Therefore, this approach is used in the milestone of pre-evaluation of the number of clusters as well as in the clustering implementation in the presented research approach (see Fig. 2).

3.4 Clustering Techniques

While some cluster approaches require a predefined k, other clustering approaches determine the number of clusters by internal evaluation (cf. Section 2.2). The clustering techniques in the current context need to be usable for textual data (Agnihotri et al. 2014; Jensi and Jiji 2014; Pons-Porrata et al. 2007) and in particular for high-dimensional communication data of electronic negotiations. Consequently, four clustering techniques are applicable and will be evaluated in our research framework in terms of their performance: k-Means, x-Means, DBSCAN and Hierarchical agglomerative clustering (see Figs. 1 and 2).

The k-Means clustering technique has most often been used in various real-world scenarios (Agnihotri et al. 2014; Jain 2010). Taking into account the described cosine similarity, k-Means clustering can be computationally efficient,

especially for sparse high-dimensional data vectors which result from the transformation of natural language and documents (Jun et al. 2014; Ravindran and Thanamani 2015). The basic idea of k-Means is to divide the underlying data objects into a predefined number of clusters where individual instances are assigned to the cluster with the closest cluster centre. The cluster centre is determined in an iterative manner until no cluster centre can be found that has a lower overall distance to the cluster instances than the current centre and thus fulfils its convergence criterion (Kassambara 2017; Khan and Ahmad 2004).

The x-Means approach evaluates different cluster distributions in the internal process by taking into account the BIC criterion. Hence, the x-Means clustering is not only able to determine the number of clusters k automatically but also to divide the underlying dataset into corresponding cluster groups (Pelleg and Moore 2000). It is thus an extension of the k-Means approach.

In comparison to other clustering algorithms, DBSCAN is better able to distinguish closely packed clusters of arbitrary shape and clusters the dataset based on the minimum level of density of data objects in the underlying data (Ikonomakis et al. 2019; Schubert et al. 2017). It is assumed that data points in the same high-density area form a cluster and different clusters are separated by low-density areas with only few instances (Xu and Tian 2015). To determine the areas, two parameters are defined, namely the radius eps of the neighbourhood and the minimum number of points MinPts in the neighbourhood (Ester et al. 1996). These quantities are determined at the beginning and are subsequently used for the entire clustering process by iteratively expanding the clustering groups (Yuan et al. 2017). This allows clusters to be found in arbitrary shapes. DBSCAN detects outliers in low-density areas and does not require the number of clusters to be predefined (Benabdellah et al. 2019; Kuwil et al. 2019).

The last clustering technique to be evaluated in this research paper is the agglomerative clustering which belongs to the hierarchical clustering methods. Hierarchical procedures enable to establish the relationships between cluster super-groups and sub-groups without pre-defining the number of clusters by using and interpreting a so-called dendrogram (Hu and Yoo 2006). A dendrogram presents a visualisation of the binary cluster divisions and enables the determination of the optimal cluster number (Forina et al. 2002; Yim and Ramdeen 2015). Agglomerative hierarchical clustering methods thus successively merge the most neighbouring pair of clusters to form a cluster hierarchy bottom up. All data objects initially form their own cluster whereupon the closest cluster pairs are sequentially generated into a common cluster until only one cluster exists at the highest level. The so-called average-link algorithm is used in our case for the composition of the clusters because it combines those cluster pairs where the average distance of all cluster members is minimal (Moseley and Wang 2017).

These described clustering techniques represent different approaches to clustering and need to be assessed for our holistic approach based on high-dimensional business communication data.

3.5 Performance Evaluation

In the last step of the research approach, a performance evaluation is carried out to evaluate the quality of determined cluster groups (see Fig. 2). This important step examines how well the data objects are represented in the assigned cluster groups or rather how well the cluster partitioning was performed by the techniques applied in the preliminary step (Palacio-Niño and Berzal 2019). In particular in high-dimensional space, performance evaluation is a major challenge and, therefore, needs to be optimised iteratively for determining the optimal number of clusters as well as the clustering techniques to be applied. This is done until a cluster split can be found that provides the best performance evaluation values (Tomašev and Radovanović 2016). To this end, external metrics are used in addition to selected internal evaluation metrics from the optimisation of k to assess the formed clusters, to be able to guarantee a holistic perspective and to provide the clearest assessment of the cluster quality. The previously described internal evaluation criteria are characterised by the sole use of internal cluster information (Liu et al. 2010) whereas external evaluation criteria compare the results of previous clustering to externally defined class labels. This shows to what extent external prediction methods are able to predict the same cluster labels, so that complementary information about cluster quality can be derived (Rokach and Maimon 2005).

Consequently, two internal and two external evaluation criteria are considered and compared within the framework of our research approach. The already presented DB index and Silhouette index are calculated for the performance evaluation of performed clustering techniques as internal evaluation metrics for the final cluster splits from the previous milestone. In addition to the internal evaluation, the balanced F-score and Cohen's Kappa are used as external performance evaluation. Since the research approach of this paper pursues the goal of evaluation-based determination of the best cluster distribution, it is necessary to determine the cluster distribution of applied techniques that maximises the predictive power within the framework of the external evaluation. The cluster labels determined in the preliminary step are assumed to be the target classes to be predicted in order to check how well the underlying data points can be traced back to the generated cluster labels using predictive forecasting techniques (Rokach and Maimon 2005).

The F-score represents the first quality measure and takes the average measures of prediction and recall for the calculation into account. Especially in the case of an unbalanced class distribution which can result from the cluster distribution in our application context, the F-score is a sensitive indicator. While an F-score value of 1 represents a very good prediction result as the best case, a value of 0 stands for a very poor result (van Rijsbergen 1979). Since a multi-class problem exists in our external evaluation context, an average overall F-score is crabwise calculated in our use case. Consequently, a corresponding F-score is calculated for each cluster label of the respective clustering approach to be predicted. The average of all F-scores is then calculated based on the previous intermediate results (Grandini et al. 2020).

In addition to the average F-score, Cohen's Kappa represents the next renowned indicator for measuring the quality of prediction that is going to be applied to multi-class problems with unbalanced class distributions. The results of the Kappa

calculation can vary between a value of −1 and a value of + 1 (Cohen 1968). A negative value indicates unusable prediction results while a value between 0.81 and 1 indicates a nearly perfect prediction result (Landis and Koch 1977; McHugh 2012). Cohen (1968) recommends Kappa values greater than 0.4 to be considered within the acceptable range and that prediction results below this threshold should not be used.

The evaluation of external performance measures is carried out by means of a predictive ensemble approach using a regression classification learner in the context of this work. This approach benefits from the support vector approach according to Vapnik (1998), whose modelling approach has successfully been used for various application in the textual high-dimensional domain (Dadgar et al. 2016; Zhang et al. 2008). Moreover, the ensemble approach focuses on a multi-class problem, which is not feasible with a standard SVM technique. Therefore, a polynomial classification using a regression learner is used to predict the clustering labels in this paper. A regression model is trained for each cluster label to be predicted. Afterwards, the individual regression models are combined into a joint classification model and the prediction performances of individual cluster labels are finally measured (Awad and Khanna 2015).

To summarise, the underlying communication data will be processed step by step through these presented milestones of the research approach, so that (1) the optimal number of clusters will be evaluated using various optimisation approaches, (2) a broad potential of clustering techniques will be applied to the underlying data, and (3) the performance of applied clustering approaches will be finally measured. These steps are of particular importance to determine the best cluster distribution and for a holistic view by including different perspectives.

4 Results

We will now report on the analytical comparison of renowned clustering techniques on high-dimensional communication data from e-negotiations in Negoisst (cf. Section 2.1). Ten negotiation experiments of several hundred participants with a total of 7026 exchanged negotiation interactions were collected between 2010 and 2016 to conduct the presented research approach. These experiments were conducted as part of university courses with students from various Bachelor's and Master's programmes worldwide.

According to the research approach, the communication data is subjected to pre-processing in the first three steps and divided into 72,826 communication units in the form of negotiation sentences. It should be noted that the incoming negotiation greeting of each negotiation message was coded by default for simplicity. Therefore, cluster sizes of greater than two would have to be considered in the case of cluster generation. On this basis, associated term document matrices (TDMs) are formed from the negotiation sentences in the next milestone. TDM represents a list of words exchanged in the negotiation sentences in a matrix; it depicts the negotiation units (the documents) to be examined as rows and the underlying terms occurring in the negotiation as dimensions (Anandarajan et al.

2019). These dimensions are efficiently compressed by iteratively conducting different dimensionality reduction algorithms (Kaya and Schoop 2020). Thereby, a TDM reduced with OS with a matrix size of [72,826 × 4841] as well as a TDM reduced with the PCA method with a size of [72,826 × 175] were found to be efficiently compressed in further studies. It should be noted that a more intensive reduction with a total of 175 dimensions (out of initially 8880 dimensions) was made by the Feature Extraction approach with PCA (PCA-dataset) compared to the feature selection approach of OS (OS-dataset) with a total of 4841 dimensions (out of initially 9661 dimensions). To prepare the datasets for cluster evaluation, both datasets are used in the last preparation step according to the underlying quantification metric with regard to the calculation of the Cosine Similarity measure. These steps are of particular importance for carrying out the subsequent evaluation milestones in an efficient manner.

The determination of the optimal number of clusters is evaluated using the four determination approaches described in the previous chapter. It should be noted that the individual optimisation methods can only provide an overall picture of the optimal number of clusters in total. Consequently, each of the solution indicators, which are presented below, should be considered as important even if no clear results can be provided in some cases due to the high-dimensionality, so that all results will be examined and discussed in a holistic manner with regard to the overall research question.

In the elbow method, an elbow point has to be determined to propose an optimal number of clusters for the underlying dataset. The examination of the course of the OS-dataset in Fig. 3 shows that no clear answer can be given as to the number of clusters by exclusively considering the elbow method in this particular case. The curve of the average centroid distance contains several ups and downs over the evaluated number of clusters. Consequently, interval containments for a k of [3;5] as well as [7;9] can be made in the first step taking these results into account. These containments enable the highlighting of the interval range where the optimal number of k could be hidden with a high probability with respect to the high-dimensional space (see Fig. 3). Therefore, other results are required for a clear statement about the optimal number of clusters for the OS-dataset.

In contrast, the results of the elbow method of the PCA-dataset show clear results. An optimal number of clusters of 5 is suggested for the PCA-dataset (see Fig. 4), since the average centroid distance examined for a reduced number of principal components shows a clearly recognisable elbow in the curve.

As a second optimisation procedure, the DB index is evaluated in an iterative manner for both datasets in the next step. As described in the previous chapter, the number of clusters that minimises the DB index is assumed to be optimal here.

Taking into account the generated DB index results for the OS-dataset, it can be observed that the DB index tends to decrease with increasing k after some highs and lows and tends towards 0. Consequently, we take the lowest point before a comparatively high slope is always to be the optimal cluster number as an approximate value. In our application case, the indicators thus refer to a cluster number of five for the OS-dataset, since the DB slope falls steadily up to a k of 5 and marks the largest rebound compared to the other slopes from a $k > 5$ (see Fig. 5).

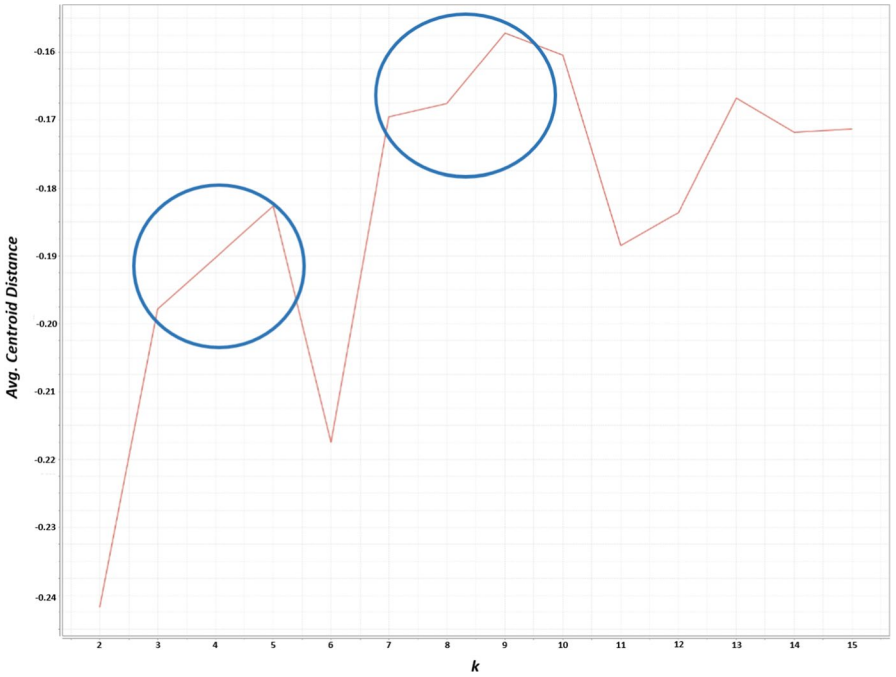

Fig. 3 Elbow method for cluster number determination based on feature selected dataset (*OS-dataset*)

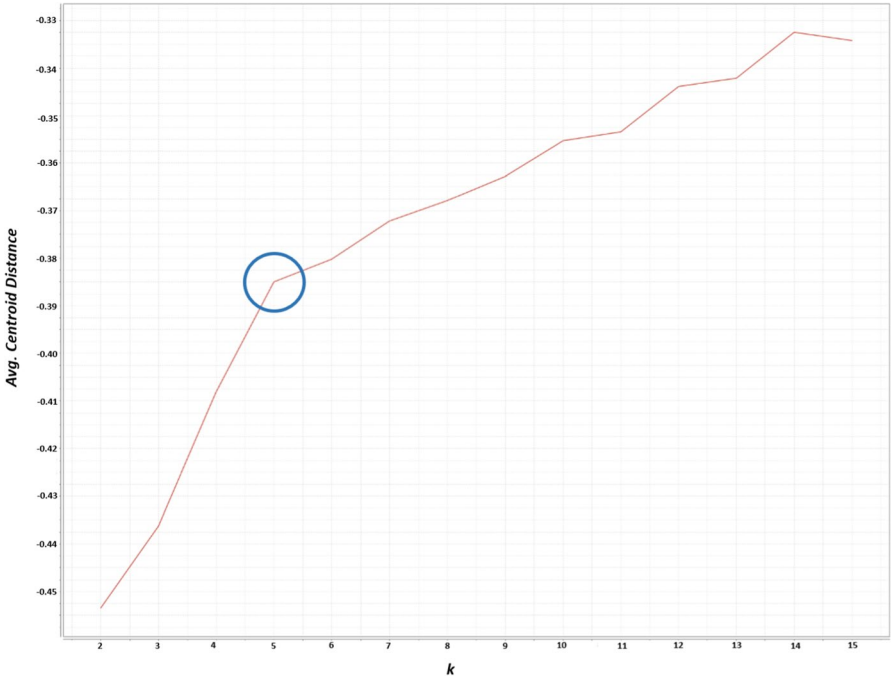

Fig. 4 Elbow method for cluster number determination based on feature extracted dataset (*PCA-dataset*)

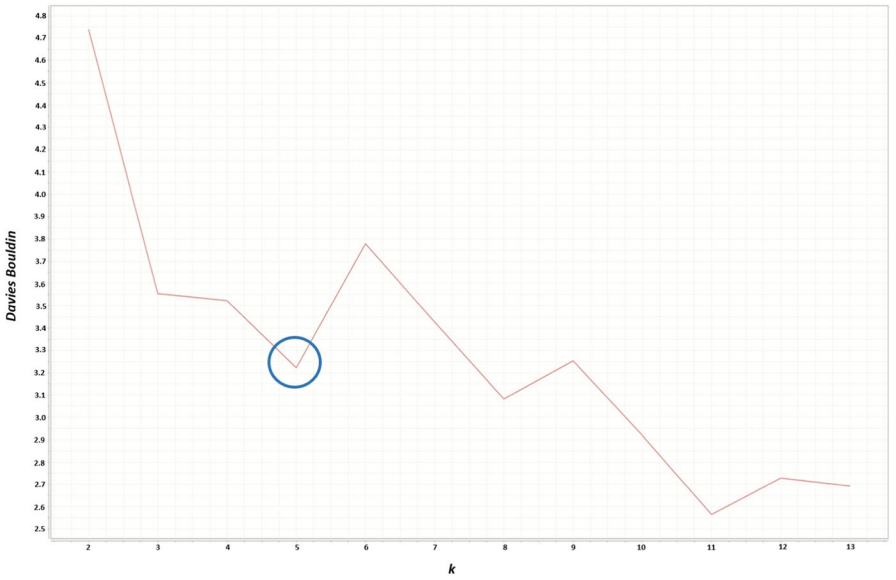

Fig. 5 Davies Bouldin evaluation based on feature selected dataset (*OS-dataset*)

A similar behaviour with ups and downs as well as a descending course of the DB index can also be observed for the evaluation of the DB index on the basis of the PCA-dataset. However, the results are less clear compared to the OS-dataset, since the course of the DB index marks several local lows without a comparatively high rebound. As a result, the DB optimisation approach is unable to suggest a number of clusters for the PCA-dataset.

In the next step, the Silhouette index is evaluated step by step for the precision of the optimal number of clusters k for the underlying PCA-dataset. According to the definition, the optimal number of clusters one that maximises the Silhouette index. As mentioned, the condition is that $k > 2$. As can be seen from the curve in Fig. 6, the curve firstly drops sharply and gradually increases before reaching its

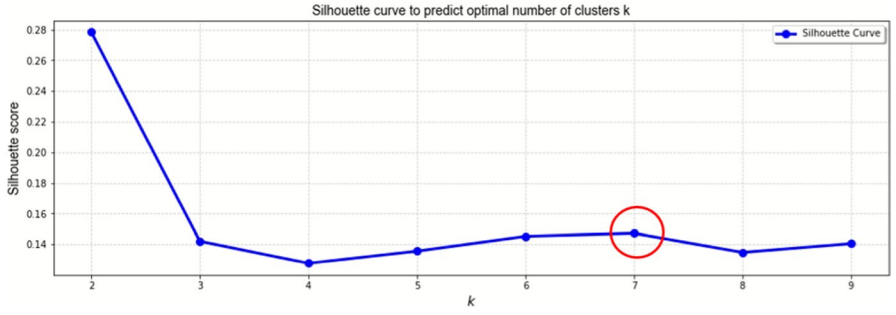

Fig. 6 Silhouette evaluation based on feature extracted dataset (*PCA-dataset*)

maximum at a k-value of 7 and subsequently dopping again, so that a value of 7 can be taken to be the optimal proposal for the PCA-dataset.

A similar curve can also be observed for the OS-dataset. However, the values for the Silhouette index are clearly higher due to the optimise selection approach with a dimensional difference of 4666 units compared to the feature extracted dataset of the PCA. After a gradual increase, the Silhouette index reaches its maximum at k = 5 and afterwards drops again for the OS-dataset (see Fig. 7). The Silhouette approach is thus able to provide clear indications regarding the determination of the optimal number of clusters for both communication datasets.

As a final approach to determine the optimal number of clusters, the x-Means approach is evaluated. Considering the underlying datasets of the OS-dataset and the PCA-dataset as well as the calculation of the BIC index, the x-Means approach is able to determine a k of 3 for both datasets in a consistent manner. Since x-Means clustering is one of those clustering methods that do not require a pre-defined number of k, this information about the number of clusters is reused exclusively for the x-Means approach in the next step.

In summary, the derived indicators of all optimisation methods uniformly indicate that a cluster splitting recommendation of k = 5 can be derived for the feature selected OS-dataset with a total of 4841 dimensions, whilst the elbow plot as well as the Silhouette index suggest an optimal cluster number of either k = 5 or k = 7 for the PCA-dataset with in total 175 dimensions. This information is used in the next step for those clustering techniques that require a pre-definition for the implementation of the respective clustering technique.

Starting with the k-Means clustering that requires the information about the proposed optimal number of clusters from the previous milestone, the k-Means approach is applied in the first step of the clustering application with an initial k = 5 and k = 7 for the PCA-dataset and subsequently with k = 5 for the OS-dataset. The cluster distributions are presented in Fig. 8. There exists one cluster group for the PCA results, which contains the same number of data objects (3389). In addition, both cluster splits contain two larger cluster groups with more than 10,000 data objects, whereby these larger clusters are further redistributed in the PCA-dataset with k = 7. In contrast to the PCA-dataset, the OS-dataset

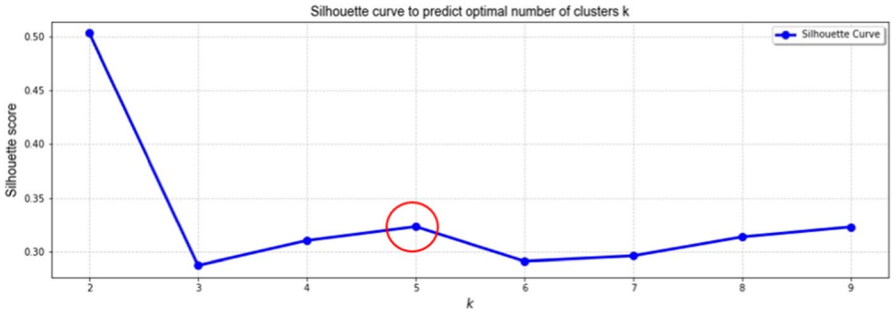

Fig. 7 Silhouette evaluation based on feature selected dataset (*OS-dataset*)

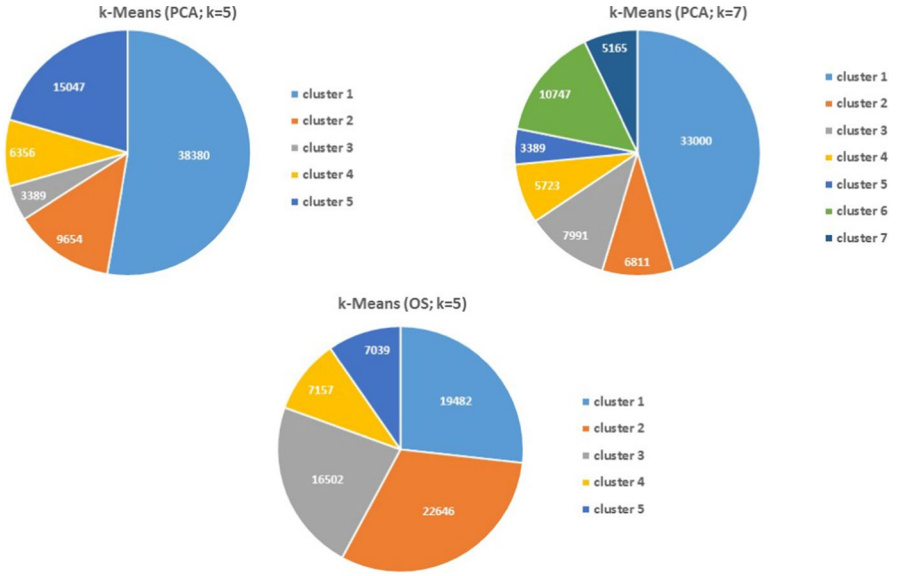

Fig. 8 Cluster partitioning for k-means clustering

with k = 5 contains three larger cluster groups. The remaining two cluster groups contain an average of 7098 data objects.

As described in the pre-evaluation to determine the optimal number of clusters, the x-Means clustering generates three clusters for both underlying datasets (see Fig. 9). It is noticeable in the internal comparison that both generated cluster groups show similar behaviour with one small exception. Both the OS-dataset as well as the PCA-dataset generate a large cluster followed by a medium-sized cluster for cluster 2. The missing difference of approximately 2000 data objects from the largest and the second largest cluster in the PCA-dataset is added to the smallest cluster from x-Means (PCA) with 10,652 data objects. Here, the smallest cluster is approximately 4000 units larger in cross-comparison with the smallest cluster of the OS-dataset with a total of 6537 data objects.

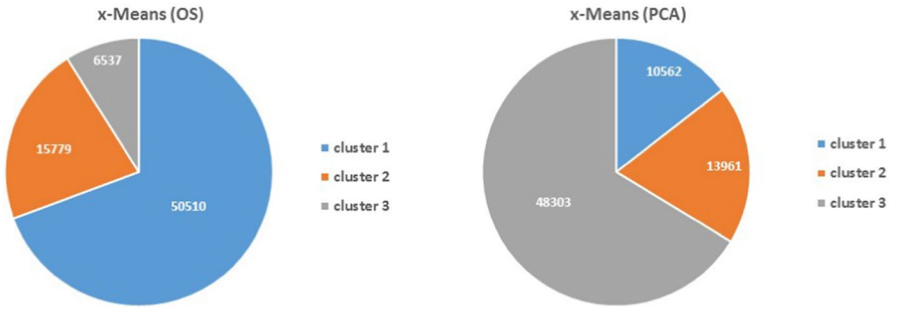

Fig. 9 Cluster partitioning for x-means clustering

The density-based approach of DBSCAN clustering does not require a predefined number of clusters either and determines the optimal number of clusters in the context of internal evaluations (Xu and Tian 2015). Based on this assumption and taking both datasets into account, DBSCAN clustering calculates a cluster distribution of four as optimal for the OS-dataset, while the same approach generates a cluster distribution of three for the compressed PCA-dataset (see Fig. 10). Looking at the determined cluster sizes, two identical cluster groups with a cluster size of 3389 data objects could be determined despite the different data. Furthermore, DBSCAN determines one very large and one smaller cluster on the basis of the PCA-dataset, while this larger cluster is split into two clusters with 46,242 and 22,646 data objects by using the OS-dataset. As the last and smallest cluster division, DBSCAN proposes a cluster with 549 data objects for the OS-dataset.

Finally, the described hierarchical agglomerative clustering is applied on the basis of the high-dimensional communication data. However, it should be noted at this point that due to the high number of underlying dimensions, no usable results could be derived for any of the two datasets. Furthermore, no statements can be made about the cluster distribution due to the dense arrangement of the hierarchical arrangements. Nevertheless, taking all cluster techniques into account, a total of 30 cluster partitions could be determined.

In order to check whether and to what extent a match exists between the individual cluster methods, all possible intersection combinations are calculated in the next step based on 30 cluster partitions.

A total of 383 intersections are checked between the respective partitions for this purpose. Table 1 of the appendix shows the average of share combinations where the intersection for both cluster partitions exceeds a 50% share of agreement. Combinations with a one-sided overlap or no intersection on both respective cluster partitions are not taken into account due to the limited informative value (see Table 1 in the appendix).

26 combinations (6.8% of all combinations) could be found which show a high degree of similarity in the internal comparison. Table 1 clearly shows, based on the individual clustering methods, that 19 methodological combinations exist that

Fig. 10 Cluster partitioning for DBSCAN clustering

have a significant share of over 75%. 9 out of 19 cluster-combinations fulfill the share constraint with at least 50% share on each cluster partition with exactly one existing cluster combination. Focusing the intra-method comparison presented in Table 1, it can be noticed that *k-Means* (*PCA; k* = 7) clustering has the largest partition-based commonality over the row-wise combinations compared to the rest of the clustering methods with a share average of 99% followed by *DBSCAN (OS)* and *k-Means (PCA; k* = *5)*. On the other hand, both x-Means approaches bear the lowest intersection with 72% (PCA) and 70% (OS) in the internal comparison. Furthermore, it is noticeable for the *k-Means (OS; k* = *5)* that no combinations could be found for four out of seven combinations due to the missing fulfillment of the minimal share condition (see Table 1).

All in all, seven different cluster separations could be achieved during the clustering process. It is still unclear which result represents the best division. Therefore, these alternatives must be evaluated in the next step using internal and external evaluation metrics. The DB index and the Silhouette index were calculated for the datasets presented in Fig. 11.

Starting with the performance evaluation results of the DB index, DBSCAN delivers the highest index value on the OS-dataset with 8.363 compared to the other DB values followed by *x-Means (OS)*'.

DB-index of 6.476. Consequently, both represent a bad seperation between clusters (Davies and Bouldin 1979). The DB index for the three k-Means variants lies between 3.456 and 4.118 and thus turns out better. The index of the second *x-Means (PCA)* lies inbetween the interval of the k-Means results with a DB value of 3.752. Nevertheless, the best DB index value (2.441) is achieved by the second DBSCAN approach based on the compressed PCA-dataset.

For the Silhouette index, the highest value is best (Rousseeuw 1987). Consequently, both k-Means approaches on PCA and the DBSCAN based on OS deliver worse Silhouette results in direct comparison based to the PCA-dataset (with k = 5 and k = 7). In contrast, the third k-Means approach achieves the best Silhouette value of 0.323 by using the OS-dataset followed by the *x-Means (OS)* approach with a Silhouette Index of 0.287 and the DBSCAN considering the PCA-dataset with a value of 0.278 (see Fig. 11). It can be noted that *DBSCAN (PCA)* performs the best cluster separation for the DB index considering these internal evaluation results. Focusing on the Silhouette index, *k-Means (OS; k* = *5)* represents the best results in the internal evaluation followed by *DBSCAN (PCA)*.

The evaluation of external performance measures is measured by the F-score and Cohen's Kappa. Those are iteratively evaluated using the regression based SVM classification approach using the underlying datasets of the multi-class clustering labels into account (see chapter 3). The results with the highest possible average F-score and Cohen's Kappa (Cohen 1968; van Rijsbergen 1979) are best. Both x-Means datasets deliver values that are clearly below average for the external evaluation metric of the average F-score. In addition, the corresponding Kappa values of both x-Means datasets are approximately 0 and thus in an unacceptable range (see Fig. 12).

In contrast, the evaluation behaviour of both DBSCAN datasets is smore positive even though both datasets deliver partially divergent results. Especially the

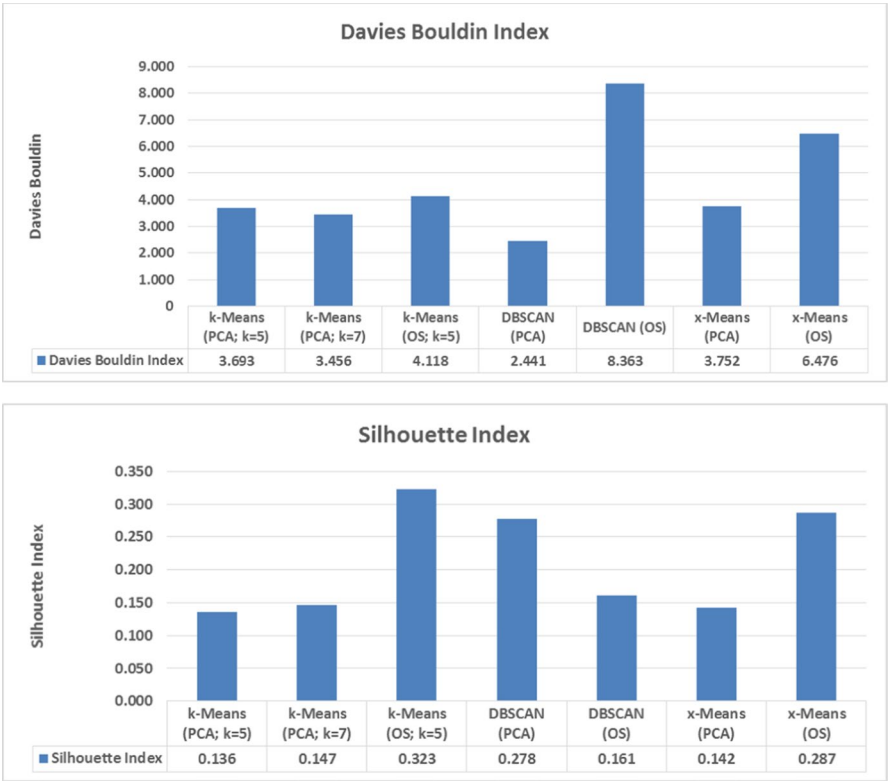

Fig. 11 Internal performance evaluation results

DBSCAN processed with the OS-dataset generates a moderately good result for the average F-score (0.698), whilst the corresponding Kappa value of 0.280 lies in an unacceptable range. Nevertheless, consistent good results could be achieved for the second DBSCAN clustering which is applied to the compressed PCA-dataset with an average F-score of 0.955 and a Kappa of 0.992.

Figure 12 shows that the k-Means techniques with k = 5 on the OS-datasets leads to an F-score of 0.604 and a Kappa of 0.432. On the other hand, k-Means on the PCA-datasets generates good to very good results for k = 5 and k = 7. The *k-Means (PCA; k = 5)* clearly performs better according to both external evaluation results due to a higher Kappa result.

In summary, three results positively stand out particularly in the context of the external evaluation: (1) *DBSCAN (PCA)*, (2) *k-Means (PCA; k = 5)* and (3) *k-Means (PCA; k = 7)*. While (1) and (2) generate similarly good results for the average F-score, (1) dominates with a very good Kappa value across all cluster separations followed by (2). As additional result, datasets (1) and (2) stand out positively with the best external evaluation performance results.

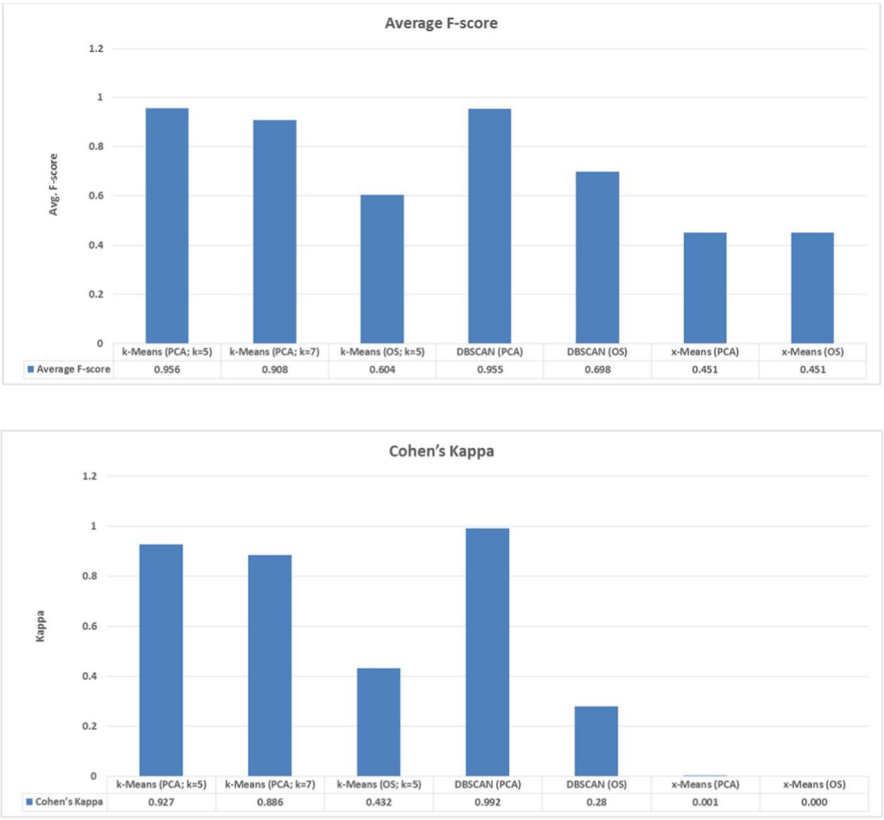

Fig. 12 External performance evaluation results

5 Discussion

As described in the previous chapters, machine-based processing reveals numerous challenges in terms of performing clustering of high-dimensional communication data. Consequently, this research paper investigates the overall research question whether and to what extent renowned clustering techniques are suitable for the recognition of groups of patterns, whether and how the respective clustering approaches differ considering high-dimensional communication data from the negotiation environment.

The presented research approach (chapter 3) and the derived results (chapter 4) show that the central research question must be answered from a holistic perspective, i.e. the evaluation milestones from the presented research approach for determining the optimal number of clusters, the active cluster evaluation, and the performance evaluation jointly contribute to answering the research questions.

The generation of a recommendation for determining the optimal number of clusters is evaluated based on the compressed PCA-dataset and the reduced OS-dataset

using four optimisation methods in the context of this research paper. It emerges that the Silhouette index and the x-Means approach using an internal BIC score are the optimisation methods that are able to derive explicit recommendations with regard to the optimal number of clusters for both underlying datasets.

However, the evaluation of the elbow method and the DB index are by no means negligible. The elbow method gives a clear recommendation on the number of clusters for the PCA-dataset (k = 5) but performs less accurately on the OS-dataset which contains significantly more dimensions. Nevertheless, information could be derived with potential optimal solutions in two intervals ([3;5] and [7;9]) for the OS-dataset. Hence, the results of the elbow method regarding the OS-dataset should be interpreted as complementary information to limit the range of optimal solutions. Especially the high-dimensionality poses the challenge that the consideration of sole procedures is not always sufficient for evaluation or optimisation problems. Therefore additional approaches must be considered to determine the solution spectrum as in the case of our research approach (Halkidi et al. 2000; Maulik and Bandyopadhyay 2002). It remains to be seen whether and to what extent precision is achieved via additional following optimisation approaches in this respect.

The DB index also generates partially approximated statements regarding the prediction of the optimal number of clusters. However, in direct comparison to the elbow method, an approximately unambiguous result exists for the OS-dataset (k = 5) despite the high-dimensional density. Surprisingly, no result determination for the PCA-dataset is possible due to the existence of several local minimums. These observations reveal that the number of high dimensions is not always the trigger for obtaining approximated results and that the optimisation method used can also play an important role in the optimisation cluster number.

In contrast, the Silhouette index provides unambiguous results indicating a clear alignment of the optimal number of clusters for both datasets. It is also noticeable that the range of values of the calculated Silhouette indices between the OS-dataset and the PCA-dataset differ significantly in the internal comparison. Higher values are for the feature selected dataset of OS. This observation suggests the data points of the cluster distribution to be further away from the neighbouring cluster groups according to the OS-dataset and thus to represent a better separation of the clusters in a direct comparison (Aranganayagi and Thangavel 2007).

The x-Means approach calculates unique results compared to the remaining optimisation methods with k = 3 in each case taking into account the BIC according to Pelleg and Moore (2000). However, the results of the x-Means lean towards smaller cluster separations and thus deviate significantly from the rest of the results.

From a meta-level perspective, two of four approaches (elbow method and Silhouette index) agree on the optimal number of clusters amounting to k = 5 based on the PCA-dataset. For the OS-dataset, two approaches (DB index and the Silhouette index) propose unique cluster separations whilst the elbow method proposes an interval delimitation. Nevertheless, it is remarkable that the interval proposal of the elbow method goes hand in hand with the results of the DB index and Silhouette index so that no contradictory predictions exist. The proposed cluster number of 5 for DB agrees with the first interval of [3;5] as well as the result of the Silhouette index of 7 with the second interval of [7;9] although two intervals could be derived

by the elbow method. The result of the x-Means evaluation in the amount of 3 also falls into the first predicted interval section despite a comparatively smaller cluster separation. In the context of cross-comparison, this finding shows that all partial results of evaluated optimisation approaches reflect important complementary information with regard to answering the overall research question in a clear way and consequently need to be investigated and discussed under a holistic magnifying glass in order to be able to determine a global solution (Singh et al. 2020). This has been performed in the context of this work.

As explained in the previous chapters, k-Means clustering is the only method that requires a pre-defined optimal number of clusters. Hence, the information from the previous milestone could be reused and the respective clusters are generated for the PCA-dataset and the OS-dataset. It should be noted with regard to the derived cluster distribution of the PCA-dataset that similar structures between the two generated cluster separations exist (see Fig. 8). This could be related to the fact that k-Means clustering generally has a high noise resistance compared to other clustering techniques (Benabdellah et al. 2019). Both cluster results contain two larger clusters and also have a cluster group that contains the identical number of objects. This cluster group could be indicative of the pre-coded standard greeting units and should be analysed in the course of further steps as with the rest of the clusters (Maqbool and Babri 2005; Muhr et al. 2010). Furthermore, it can be stated for the results of the PCA-dataset that there might be further redistribution of data objects from the two large cluster groups leading to an increase in the number of clusters from 5 to 7. The majority of added data objects originate from the two large cluster groups concerning the quantity of generated clusters which have significantly fewer data objects than before. Additionally, a slightly different cluster structure is achieved this time in the third k-Means application based on the OS-dataset compared to the PCA-dataset despite using the same method. While both PCA-datasets are able to determine two larger cluster groups, the k-Means suggests three larger cluster groups based on the OS-dataset. In addition, the unique cluster group with 3389 data objects could not be determined using the OS-dataset. A potential reason for this discrepancy might be the differing dimensionality reduction algorithms leading to a significantly higher dimensional density for OS compared to PCA. This phenomenon can influence the Machine Learning procedures of clustering in a significant way (Fan and Li 2006; Shah and Patel 2016). For all three results from Fig. 8, the generated cluster size should not be disregarded as the k-Means generally tends to similarly sized cluster groups with the exception of a few larger cluster groups according to our results as well as the large underlying datasets. The k-Means approach thus has a good scalability and performs best with large datasets (Kuwil et al. 2019).

DBSCAN determines the optimal number of clusters in an automated way by taking the data density into account (Duan et al. 2007). We could show that DBSCAN can determine smaller cluster groups for both datasets compared to the other clustering techniques (see Fig. 10). Nevertheless, two large cluster groups are determined for the OS-dataset, similar to the k-Means approach. It should be mentioned that the scalability of DBSCAN increases with increasing number of instances so that it is roughly on the same level as k-Means (Mavridis et al. 2013; Xu and Tian 2015). However, the results of the PCA-dataset differ as it does not decide for a further

split of the larger cluster with 69,217 data objects. The phenomenon of determining smaller cluster groups next to very large cluster groups might be due to the fact that DBSCAN is better able to distinguish tightly packed clusters as it does not use a distance measure (Ikonomakis et al. 2019). In addition, the density-based approach could result in the data objects not being fully deterministic at the boundary points which can be reached by multiple clusters. Furthermore, DBSCAN might have difficulties when the density of different clusters varies since the density threshold represents a global parameter (Kuwil et al. 2019). This may result in small clusters being identified right next to larger clusters that should be included in the adjoining larger cluster. Nevertheless, there should always be a scientific basis before deciding to reintegrate certain cluster groups. This is the reason why we have analysed and compared several cluster techniques.

As presented in the results chapter, the x-Means clustering using the internal BIC evaluation criterion generates three cluster groups for each of the two datasets. Thus, the results of the x-Means refer to the same internally determined optimal number of clusters as in the result of the PCA-dataset from DBSCAN. Nevertheless, an in-depth analysis of the cluster separations of both techniques reveals an elementary difference in terms of cluster size. While the density-based clustering tends to smaller cluster sizes, derived cluster groups of the x-Means are significantly larger with the exception of only one group.

Finally, the agglomerative hierarchical approach could not provide usable results due to the large number of data objects of considered communication units in the context of our application. Consequently, the general disadvantage of hierarchical clustering is that it tends to produce very small clusters that may contain outliers and thus manipulate the cluster separation (Cetinkaya et al. 2015; Hu et al. 2018). This information provides an important rationale regarding the way hierarchical clustering performance is affected when using noisy data in high-dimensional space.

Across all results, it can be observed that the cluster sizes differ between three and seven cluster separations considering the results from the pre-evaluation to determine the optimal number as well as internal evaluation metrics of the x-Means and DBSCAN. Although different pre-processing structures exist for the OS-dataset and the PCA-dataset, individual clustering techniques are able to show structural similarities in addition to structural differences. As explained in the previous chapter, the average share of cluster combinations in Table 1 of the appendix clearly show the existence of a series of patterns with identical as well as similar cluster mappings despite the use of different pre-processing and clustering techniques (see Table 1). This finding highlights that identical or rather highly similar structures exist in the underlying communication data.

All evaluated clustering techniques are able to operate on the high-dimensional communication data and generate specific cluster separations depending on the underlying algorithm with one exception regarding the hierarchical clustering approach.

The final performance evaluation touches on the aspect of measuring the quality of determined clusters and thus makes a contribution to answer the second part of the research question (i.e. how well the cluster results reflect the considered communication data). As discussed previously, external measures are also used in addition

to the established internal evaluation measures. Especially in the high-dimensional space, where the large number of dimensions and data objects can lead to methodological challenges, this combined approach is helpful to derive the most reliable statements about the quality of generated cluster groups (Liu et al. 2005; Maulik and Bandyopadhyay 2002).

Hence, the combination of both measures shows that the generated cluster results may be different for internal and external evaluation measures. Whilst both lead to similar cluster splits according to our results so that obtained results are mutually supportive, other splits fail to validate either on the internal or external evaluation side. The *k-Means* (*OS; k = 5*) achieves the best Silhouette index score and produces a performance with an average F-score and Kappa values when only external evaluation measures are considered. On the other hand, the Cohen's Kappa as well as the F-score unanimously suggest the *k-Means* (*PCA; k = 7*) as the third best cluster split. Nevertheless, this result remains unsupported, this time on the internal evaluation side. These findings show the fundamental importance of integrating both internal and external evaluation metrics as complementary information (Rendón et al. 2011). Nevertheless, these results should be maintained for now, especially if the cluster breakdowns are supported by either the internal or the external evaluation measures. For example, there is clear agreement on the positive effect of cluster separation in the case of the *k-Means* (*PCA; k = 7*) but not for the *k-Means* (*OS; k = 5*) because of the DB index.

The overall performance results show that the *DBSCAN (PCA)* is suggested as a good cluster split by both internal and both external performance measures. The corresponding values always scores among the top two results. Furthermore, the compressed PCA-dataset with the *k-Means* (*PCA; k = 5*) is recommended as a valid cluster split by focusing on the external evaluation measures. In contrast, cluster results that used the OS-dataset perform only average from an external performance perspective. The high number of dimensions in the OS-dataset might have negatively influenced the external evaluation and increased the complexity in predicting the correct cluster labels (Tibshirani and Walther 2005). On the other hand, the PCA approach significantly reduced the number of dimensions by combining existing dimensions into coherent dimensions based on their similar structure. This ensures the most effective retention of information density despite a reduction in quantity, which benefits Machine Learning methods (Abdi and Williams 2010; Wold et al. 1987).

In addition to valid clustering results, some results perform below average. Both x-Means results show poor to moderate performance results according to Cohen's Kappa as well as the average F-score followed by *DBSCAN* (*OS*).

6 Conclusion and Outlook

The processing and evaluation of unstructured textual communication data is challenging for pattern recognition due to the missing structure and the high number of dimensions (Bonev et al. 2008; Donoho 2000). We presented a structured approach for the recognition of groups of patterns for electronic negotiation communication

data. In particular, a research approach with three evaluation milestones for the determination of the optimal number of clusters, the application potential of selected clustering techniques, and the subsequent performance evaluation was elaborated and experimentally evaluated to investigate the application potential on pre-processed negotiation communication data of the Negoisst system. Taking into account the results and analytical discussion of all integrated approaches, it was shown that certain cluster separations could be eliminated with a clear answer, whereas further cluster separations could either be recommended by exclusively internal (*k-Means* (*OS; k=5*)), external (*k-Means* (*PCA; k=5 and k=7*)) or both (*DBSCAN* (*PCA*)) performance evaluation measures. The results show that each of the described evaluation milestones contain elementary approaches that provide usable insights for determining cluster groups. The holistic perspective of assessing all approaches with different data sets reveals new possibilities and insights regarding the reactionary behaviour of the applied methods.

When it comes to generalisability, it should be noted that the data basis consists of sentence units of entire negotiation communications. Consequently, no phase-related splitting of communication units was conducted which divides the entire negotiation e.g. into theoretically motivated negotiation segments (Adair and Brett 2004; Weingart et al. 2004). Considering more abstract levels of aggregation might imply further potential concerning the formation of further and possibly different cluster groups instead of underlying the whole negotiation course. Furthermore, the developed research approach naturally reveals potential for the evaluation of further optimisation techniques. Additional clustering approaches and internal and external performance evaluation measures can be integrated to extend the scope of the current research.

The findings of this research paper offer numerous possibilities regarding a future research potential. In addition to determining further cluster groups at different aggregation levels of negotiation communication, further research steps could include a detailed content-related analysis of individual cluster groups which perform best, e.g. in terms of the evaluated performance. This would provide important insights regarding the interpretability of derived pattern groups and thus reveal what kind of patterns are hidden behind individual clusters (Role and Nadif 2014). These interpreted cluster groups could reveal further specifics of bargaining behaviour and thus provide a broader picture in a combined study of different aggregation levels. Previous work on determining behavioural negotiation patterns (Pesendorfer et al. 2007; Sokolova et al. 2004) identified indicators for strategic orientations, concession behaviour, information disclosure, emotions etc. (McGinn et al. 2003; Van Kleef et al. 2004; Vetschera 2016). These individual behavioural components can collectively provide key cues for particular strategic and behavioural orientations through a systematic linkage (Weingart and Olekalns 2004). An automated content-based investigation of negotiation sentences based on our work would serve this goal and provide important indicators for behavioural orientations. The methodological spectrum of Machine Learning offers various possibilities such as topic modelling techniques, association rule analyses, and other predictive methods to describe detected clusters considering the underlying data objects (Lee and Lee 2005; Xie and Xing 2013).

In addition to the consideration of further levels of aggregation as well as the exploration of the descriptive labelling potential of detected clusters, the dynamic perspective of negotiation processes could be integrated in future studies. Especially the bilateral exchange of negotiation messages in episodic negotiation phases, the change of negotiation strategies in form of sequences could provide further exciting indicators for clustering (Druckman 2001; Olekalns et al. 2003; Olekalns and Weingart 2008). Automated precoding could thus be used to derive further metrics which could then be incorporated into the algorithmic processing of pattern recognition, either separately or as part of a holistic perspective. Whilst complete negotiation sentence units have been considered for discussing the pattern recognition potential of electronic negotiation messages and methodological challenges, meta-information such as the sequence of preceding or following sentences, possible information about the message type, and timestamps of sent messages could be considered in addition to previously dynamic metrics. This complementary information might support an extended value-added contribution to the automated determination of negotiation behaviour.

The research field of pattern recognition in natural communication language is broad and offers numerous descriptive as well as predictive possibilities for structuring and interpreting unstructured natural language in terms of semantics in a systematic way. Especially for negotiation communication, further investigation can provide an important value-adding contribution to the efficient as well as effective conduct of business negotiations (Donohue and Roberto 1996; Hargie and Dickson 2004). The presented evaluation approaches can be applied to other areas for the recognition of pattern groups in natural language. This allows for the automated recognition of important structures in big datasets. It is, therefore, even more important to take up the challenges of the high-dimensional data space and to work on related problems from an analytical perspective. The world of Machine Learning remains will continue to make an important contribution to the development of data-driven research in future.

Appendix

See Table 1.

Table 1 Average of the largest sum of cluster shares between the clustering methods (at least 50% share on both respective cluster partitions)

	k-Means (PCA; k = 7)	k-Means (PCA; k = 5)	k-Means (OS; k = 5)	DBSCAN (PCA)	DBSCAN (OS)	x-Means (PCA)	x-Means (OS)
k-Means (PCA; k = 7)	**1**	0.97	–	1[a]	1[a]	0.99	–
k-Means (PCA; k = 5)	0.84	**1**	–	1[a]	0.99	0.96	0.69[a]
k-Means (OS; k = 5)	–	–	**1**	–	1[a]	–	0.67
DBSCAN (PCA)	1[a]	0.78	–	**1**	0.84	1[a]	0.68
DBSCAN (OS)	1[a]	1[a]	1[a]	1	**1**	0.67[a]	–
x-Means (PCA)	0.60	0.67	–	1[a]	0.64[a]	**1**	0.69[a]
x-Means (OS)	–	0.52[a]	0.71	0.93	–	0.66[a]	**1**

[a]Only one existing cluster combination fulfilling the minimal share constraint

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abdi H, Williams LJ (2010) Principal component analysis. Wiley Interdiscip Rev Comput Stat 2:433–459. https://doi.org/10.1002/wics.101

Abualigah LM, Khader AT, Al-Betar MA (2016) Multi-objectives-based text clustering technique using K-mean algorithm. In: 7th international conference on computer science and information technology (CSIT), pp 1–6

Adair WL, Brett JM (2004) Culture and negotiation processes. In: Gelfand MJ, Brett JM (eds) The handbook of negotiation and culture. Stanford University Press, pp. 158–176

Adair WL, Brett JM (2005) The negotiation dance: time, culture, and behavioral sequences in negotiation. Organ Sci 16:33–51. https://doi.org/10.1287/orsc.1040.0102

Agnihotri D, Verma K, Tripathi P (2014) Pattern and cluster mining on text data. In: Fourth international conference on communication systems and network technologies, pp 428–432

Allahyari M, Pouriyeh S, Assefi M, Safaei S, Trippe ED, Gutierrez JB, Kochut K (2017) A brief survey of text mining: classification, clustering and extraction techniques. arXiv:1707.02919

Anandarajan M, Hill C, Nolan T (2019) Term-document representation. In: Anandarajan M, Hill C, Nolan T (eds) Practical text analytics. Springer, Cham, pp 61–73

Aranganayagi S, Thangavel K (2007) Clustering categorical data using silhouette coefficient as a relocating measure. In: International conference on computational intelligence and multimedia applications (ICCIMA 2007), vol 2, pp 3–17. https://doi.org/10.1109/ICCIMA.2007.328

Awad M, Khanna R (2015) Support vector regression. In: Awad M, Khanna R (eds) Efficient learning machines. Apress, Berkeley, pp 67–80

Bafna P, Pramod D, Vaidya A (2016) Document clustering: TF-IDF approach. In: International conference on electrical, electronics, and optimization techniques (ICEEOT), pp 61–66

Benabdellah AC, Benghabrit A, Bouhaddou I (2019) A survey of clustering algorithms for an industrial context. Procedia Comput Sci 148:291–302. https://doi.org/10.1016/j.procs.2019.01.022

Bhagat A, Kshirsagar N, Khodke P, Dongre K, Ali S (2016) Penalty parameter selection for hierarchical data stream clustering. Procedia Comput Sci 79:24–31. https://doi.org/10.1016/j.procs.2016.03.005

Bholowalia P, Kumar A (2014) EBK-means: a clustering technique based on elbow method and k-means in WSN. Int J Comput Appl 105:9. https://doi.org/10.5120/18405-9674

Bichler M, Kersten G, Strecker S (2003) Towards a structured design of electronic negotiations. Group Decis Negot 12:311–335. https://doi.org/10.1023/A:1024867820235

Bonev B, Escolano F, Cazorla M (2008) Feature selection, mutual information, and the classification of high-dimensional patterns. Pattern Anal Appl 11:309–319. https://doi.org/10.1007/s10044-008-0107-0

Cetinkaya S, Basaraner M, Burghardt D (2015) Proximity-based grouping of buildings in urban blocks: a comparison of four algorithms. Geocarto Int 30:618–632. https://doi.org/10.1080/10106049.2014.925002

Chen Y, Tang S, Bouguila N, Wang C, Du J, Li H (2018) A fast clustering algorithm based on pruning unnecessary distance computations in DBSCAN for high-dimensional data. Pattern Recogn 83:375–387. https://doi.org/10.1016/j.patcog.2018.05.030

Cohen J (1968) Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70:213–220. https://doi.org/10.1037/h0026256

Croson RT (1999) Look at me when you say that: an electronic negotiation simulation. Simul Gaming 30:23–37. https://doi.org/10.1177/104687819903000105

Dadgar SMH, Araghi MS, Farahani MM (2016) A novel text mining approach based on TF-IDF and support vector machine for news classification. In: IEEE international conference on engineering and technology (ICETECH), pp 112–116
Das TK, Kumar PM (2013) Big data analytics: a framework for unstructured data analysis. Int J Eng Sci Technol 5:153–156
Davidson I, Ravi SS (2005) Agglomerative hierarchical clustering with constraints: theoretical and empirical results. In: European conference on principles of data mining and knowledge discovery. Springer, Berlin, Heidelberg, pp 59–70
Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1:224–227
Dharmarajan A, Velmurugan T (2013) Applications of partition based clustering algorithms: a survey. In: IEEE international conference on computational intelligence and computing research, pp 1–5
Donoho DL (2000) High-dimensional data analysis: the curses and blessings of dimensionality. AMS Math Chall Lect 1:32
Donohue WA, Roberto AJ (1996) An empirical examination of three models of integrative and distributive bargaining. Int J Confl Manag 7:209–229. https://doi.org/10.1108/eb022782
Druckman D (2001) Turning points in international negotiation: a comparative analysis. J Conf Resolut 45:519–544
Duan L, Xu L, Guo F, Lee J, Yan B (2007) A local-density based spatial clustering algorithm with noise. Inf Syst 32:978–986. https://doi.org/10.1016/j.is.2006.10.006
Erk K (2012) Vector space models of word meaning and phrase meaning: a survey. Lang Linguist Compass 6:635–653. https://doi.org/10.1002/lnco.362
Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd 96:226–231
Fan J, Li R (2006) Statistical challenges with high dimensionality: feature selection in knowledge discovery. arXiv preprint math/0602133
Feldman R, Sanger J (2007) The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, Cambridge
Forina M, Armanino C, Raggio V (2002) Clustering with dendrograms on interpretation variables. Anal Chim Acta 454:13–19. https://doi.org/10.1016/S0003-2670(01)01517-3
Frades I, Matthiesen R (2010) Overview on techniques in cluster analysis. Bioinformatics methods in clinical research. Humana Press, Totowa, pp 81–107
Fred AL, Leitao JM (2000) Partitional vs hierarchical clustering using a minimum grammar complexity approach. Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR). Springer, Berlin, Heidelberg, pp 193–202
Gan G, Ma C, Wu J (2007) Data clustering: theory, algorithms, and applications. Soc Ind Appl Math. https://doi.org/10.1137/1.9780898718348
Gheyas IA, Smith LS (2010) Feature subset selection in large dimensionality domains. Pattern Recogn 43:5–13
Grandini M, Bagli E, Visani G (2020) Metrics for multi-class classification: an overview. arXiv:2008.05756
Gunawan D, Sembiring CA, Budiman MA (2018) The implementation of cosine similarity to calculate text relevance between two documents. J Phys Conf Ser IOP Publ 978:1–6
Habermas J (1981) Theorie des kommunikativen Handelns. Suhrkamp Verlag, Berlin
Halkidi M, Vazirgiannis M, Batistakis Y (2000) Quality scheme assessment in the clustering process. In: European conference on principles of data mining and knowledge discovery. Springer, Berlin, Heidelberg, pp 265–276
Hargie O, Dickson D (2004) Skilled interpersonal communication: research, theory and practice, 4th edn. Routledge, London
Hargie O (2010) Skilled interpersonal communication: research, theory and practice, 5th edn. Routledge. https://doi.org/10.4324/9780203833919
Hu X, Yoo I (2006) A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE. In: Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries (JCDL'06), pp 220–229
Hu CW, Li H, Qutub AA (2018) Shrinkage clustering: a fast and size-constrained clustering algorithm for biomedical applications. BMC Bioinform. https://doi.org/10.1186/s12859-018-2022-8

Huang A (2008) Similarity measures for text document clustering. In: Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand vol 4, pp 9–56

Ikonomakis EK, Spyrou GM, Vrahatis MN (2019) Content driven clustering algorithm combining density and distance functions. Pattern Recogn 87:190–202. https://doi.org/10.1016/j.patcog.2018.10.007

Irani J, Pise N, Phatak M (2016) Clustering techniques and the similarity measures used in clustering: a survey. Int J Comput Appl 134:9–14. https://doi.org/10.5120/ijca2016907841

Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recogn Lett 31:651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jensi R, Jiji DGW (2014) A survey on optimization approaches to text document clustering. arXiv:1401.2229

Jun S, Park SS, Jang DS (2014) Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Syst Appl 41:3204–3212. https://doi.org/10.1016/j.eswa.2013.11.018

Kadhim AI, Cheah YN, Ahamed NH (2014) Text document preprocessing and dimension reduction techniques for text document clustering. In: IEEE 4th international conference on artificial intelligence with applications in engineering and technology, pp 69–73

Kassambara A (2017) Practical guide to cluster analysis in R: Unsupervised machine learning Sthda

Kaya MF, Schoop M (2020) Maintenance of data richness in business communication data. In: Proceedings of the 28th European conference on information systems (ECIS), an online AIS conference

Khalid S, Khalil T, Nasreen S (2014). A survey of feature selection and feature extraction techniques in machine learning. In: IEEE science and information conference, pp 372–378

Khan SS, Ahmad A (2004) Cluster center initialization algorithm for K-means clustering. Pattern Recogn Lett 25:1293–1302. https://doi.org/10.1016/j.patrec.2004.04.007

Kodinariya TM, Makwana PR (2013) Review on determining number of cluster in K-means clustering. Int J 1:90–95

Kriegel HP, Kröger P, Sander J, Zimek A (2011) Density-based clustering. Wiley Interdiscip Rev Data Min Knowl Discov 1:231–240. https://doi.org/10.1002/widm.30

Kriegel HP, Pfeifle M (2005) Density-based clustering of uncertain data. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, pp 672–677

Kumar AC (2009) Analysis of unsupervised dimensionality reduction techniques. Comput Sci Inf Syst 6:217–227. https://doi.org/10.2298/CSIS0902217K

Kuwil FH, Shaar F, Topcu AE, Murtagh F (2019) A new data clustering algorithm based on critical distance methodology. Expert Syst Appl 129:296–310. https://doi.org/10.1016/j.eswa.2019.03.051

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174. https://doi.org/10.2307/2529310

Lee J, Lee D (2005) An improved cluster labeling method for support vector clustering. IEEE Trans Pattern Anal Mach Intell 27:461–464. https://doi.org/10.1109/TPAMI.2005.47

Lewicki RJ, Barry B, Saunders DM (2016) Essentials of negotiation. McGraw-Hill, New York

Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H (2017) Feature selection: a data perspective. ACM Comput Surv (CSUR) 50:1–45

Liu L, Kang J, Yu J, Wang Z (2005) A comparative study on unsupervised feature selection methods for text clustering. In: IEEE international conference on natural language processing and knowledge engineering, pp 597–601

Liu Y, Li Z, Xiong H, Gao X, Wu J (2010) Understanding of internal clustering validation measures. In: IEEE international conference on data mining, pp 911–916

Maldonado S, Weber R, Famili F (2014) Feature selection for high-dimensional class-imbalanced data sets using support vector machines. Inf Sci 286:228–246

Maqbool O, Babri HA (2005) Interpreting clustering results through cluster labeling. In: Proceedings of the IEEE symposium on emerging technologies, pp 429–434

Maulik U, Bandyopadhyay S (2002) Performance evaluation of some clustering algorithms and validity indices. IEEE Trans Pattern Anal Mach Intell 24:1650–1654. https://doi.org/10.1109/TPAMI.2002.1114856

Mavridis L, Nath N, Mitchell JB (2013) PFClust: a novel parameter free clustering algorithm. BMC Bioinform 14:213. https://doi.org/10.1186/1471-2105-14-213

McGinn KL, Thompson L, Bazerman MH (2003) Dyadic processes of disclosure and reciprocity in bargaining with communication. J Behav Decis Mak 16:17–34

McHugh ML (2012) Interrater reliability: the kappa statistic. Biochemia Medica: Biochemia Medica 22:276–282
Mirkin B (2012) Clustering: a data recovery approach. CRC Press, London
Morris C (1971) Writings of the general theory of signs. Mouton, The Hague
Moseley B, Wang J (2017) Approximation bounds for hierarchical clustering: average linkage, bisecting k-means, and local search. In: Advances in neural information processing systems, pp 3094–3103
Muflikhah L, Baharudin B (2009) Document clustering using concept space and cosine similarity measurement. IEEE Int Conf Comput Technol Dev 1:58–62. https://doi.org/10.1109/ICCTD.2009.206
Muhr M, Kern R, Granitzer M (2010) Analysis of structural relationships for hierarchical cluster labeling. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, pp 178–185
Munková D, Munk M, Vozár M (2013) Data pre-processing evaluation for text mining: transaction/sequence model. Procedia Comput Sci 18:1198–1207. https://doi.org/10.1016/j.procs.2013.05.286
Munot N, Govilkar SS (2014) Comparative study of text summarization methods. Int J Comput Appl 102:33–37
Myers MT, Myers GE (1982) Managing by communication—an organizational approach. McGraw-Hill Book Company, New York
Olekalns M, Weingart LR (2008) Emergent negotiations: Stability and shifts in negotiation dynamics. Negot Confl Manag Res 1:135–160
Olekalns M, Brett JM, Weingart LR (2003) Phases, transitions and interruptions: modeling processes in multi-party negotiations. Int Jo Confl Manag 14:191–211
Palacio-Niño JO, Berzal F (2019) Evaluation metrics for unsupervised learning algorithms. arXiv:1905.05667
Pelleg D, Moore AW (2000) X-means: extending K-means with efficient estimation of the number of clusters. In: Proceedings of the seventeenth international conference on machine learning. Morgan Kaufmann Publishers Inc, pp 727–734
Pesendorfer EM, Graf A, Koeszegi ST (2007) Relationship in electronic negotiations: tracking behavior over time. J Bus Econ 77:1315–1338
Petrovic S (2006) A comparison between the silhouette index and the Davies–Bouldin index in labelling ids clusters. In: Proceedings of the 11th Nordic workshop of secure IT systems, pp 53–64
Pons-Porrata A, Berlanga-Llavori R, Ruiz-Shulcloper J (2007) Topic discovery based on text mining techniques. Inf Process Manag 43:752–768. https://doi.org/10.1016/j.ipm.2006.06.001
Purdy JM, Nye P, Balakrishnan PV (2000) The impact of communication media on negotiation outcomes. Int J Confl Manag 11:162–187. https://doi.org/10.1108/eb022839
Putnam LL (2010) Communication as changing the negotiation game. J Appl Commun Res 38:325–335. https://doi.org/10.1080/00909882.2010.513999
Putnam LL, Roloff ME (1992) Communication and negotiation. Sage, London
Qu Y, Ostrouchov G, Samatova N, Geist A (2002) Principal component analysis for dimension reduction in massive distributed data sets. Proc IEEE Int Conf Data Min (ICDM) 1318:1–12
Rana MMR, Afrin R, Rahman MA, Haque A, Rahman MA (2019) Concept extraction from ambiguous text document using K-means. Int Res J Eng Technol (IRJET) 6:5317–5330
Ravindran RM, Thanamani AS (2015) K-means document clustering using vector space model. Bonfring Int J Data Min 5:10–14. https://doi.org/10.9756/BIJDM.8076
Ray S, Turi RH (1999) Determination of number of clusters in k-means clustering and application in colour image segmentation. In: Proceedings of the 4th international conference on advances in pattern recognition and digital techniques, pp 137–143
Rendón E, Abundez I, Arizmendi A, Quiroz EM (2011) Internal versus external cluster validation indexes. Int J Comput Commun 5:27–34
Renganathan V (2017) Text mining in biomedical domain with emphasis on document clustering. Healthc Inf Res 23:141–146
Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ (2006) Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. J Math Model Algorithms 5:475–504. https://doi.org/10.1007/s10852-005-9022-1
Rogers EM, Rekha AR (1976) Communication in organizations. Free Press, New York
Rokach L, Maimon O (2005) Clustering methods. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook. Springer, Boston, pp 321–352
Role F, Nadif M (2014) Beyond cluster labeling: semantic interpretation of clusters' contents using a graph representation. Knowl Based Syst 56:141–155. https://doi.org/10.1016/j.knosys.2013.11.005

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
Santhisree K, Damodaram A (2011) SSM-DBSCAN and SSM-OPTICS: incorporating a new similarity measure for density based clustering of web usage data. Int J Comput Sci Eng 3:3170–3184
Saxena A, Prasad M, Gupta A, Bharill N, Patel OP, Tiwari A, Lin CT (2017) A review of clustering techniques and developments. Neurocomputing 267:664–681
Schoop M (2004) The worlds of negotiation. In: Proceedings of the 9th international working conference of the language-action perspective on communication modelling, LAP, pp 179–196
Schoop M (2010) Support of complex electronic negotiations. In: Marc Kilgour D, Eden C (eds) Handbook of group decision and negotiation. Springer, Dordrecht, pp 409–423
Schoop M (2020) Negoisst: complex digital negotiation support. In: Kilgour DM, Eden C (eds) Handbook of group decision and negotiation. Springer, Cham. https://doi.org/10.1007/978-3-030-12051-1_24-1
Schoop M (2021) Negotiation communication revisited. Cent Eur J Oper Res. https://doi.org/10.1007/s10100-020-00730-5
Schoop M, Jertila A, List T (2003) Negoisst: a negotiation support system for electronic business-to-business negotiations in e-commerce. Data Knowl Eng 47:371–401. https://doi.org/10.1016/S0169-023X(03)00065-X
Schoop M, van Amelsvoort M, Gettinger J, Koerner M, Koeszegi ST, van der Wijst P (2014) The interplay of communication and decisions in electronic negotiations: Communicative decisions or decisive communication? Group Decis Negot 23:167–192. https://doi.org/10.1007/s10726-013-9357-3
Schubert E, Sander J, Ester M, Kriegel HP, Xu X (2017) DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Trans Database Syst (TODS) 42:1–21
Searle JR (1969) Speech acts: an essay in the philosophy of language. Cambridge University Press, Cambridge
Shah FP, Patel V (2016) A review on feature selection and feature extraction for text classification. In: IEEE international conference on wireless communications, signal processing and networking (WiSPNET), pp 2264–2268
Shah N, Mahajan S (2012) Document clustering: a detailed review. Int J Appl Inf Syst 4:30–38. https://doi.org/10.5120/ijais12-450691
Shehata S, Karray F, Kamel M (2006) Enhancing text clustering using concept-based mining model. In: IEEE sixth international conference on data mining (ICDM'06), pp 1043–1048
Silitonga P (2017) Clustering of patient disease data by using K-means clustering. Int J Comput Sci Inf Secur (IJCSIS) 15:219–221
Singh AK, Mittal S, Malhotra P, Srivastava YV (2020) Clustering evaluation by Davies–Bouldin Index (DBI) in cereal data using K-means. In: IEEE fourth international conference on computing methodologies and communication (ICCMC), pp 306–310
Sokolova M, Nastase V, Szpakowicz S (2004) Language in electronic negotiations: patterns in completed and uncompleted negotiations. In: Natural language processing (proceedings of 3rd international conference on natural language processing (ICON'2004)), pp 142–151
Swarndeep Saket J, Pandya S (2016) An overview of partitioning algorithms in clustering techniques. Int J Adv Res Comput Eng Technol (IJARCET) 5:1943–1946
Syakur MA, Khotimah BK, Rochman EMS, Satoto BD (2018) Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conf Ser Mater Sci Eng IOP Pub 336:012017. https://doi.org/10.1088/1757-899X/336/1/012017
Tibshirani R, Walther G (2005) Cluster validation by prediction strength. J Comput Graph Stat 14:511–528. https://doi.org/10.1198/106186005X59243
Tomašev N, Radovanović M (2016) Clustering evaluation in high-dimensional data. In: Celebi M, Aydin K (eds) Unsupervised learning algorithms. Springer, Cham, pp 71–107. https://doi.org/10.1007/978-3-319-24211-8_4
Tran TN, Wehrens R, Buydens LM (2006) KNN-kernel density-based clustering for high-dimensional multivariate data. Comput Stat Data Anal 51:513–525. https://doi.org/10.1016/j.csda.2005.10.001
Tutzauer F (1992) The communication of offers in dyadic bargaining. In: Putnam L, Roloff M (eds) Communication and negotiation. Sage, Newbury Park, pp 67–82
Van Kleef GA, De Dreu CK, Manstead AS (2004) The interpersonal effects of emotions in negotiations: a motivated information processing approach. J Pers Soc Psychol 87:510–528
van Rijsbergen CJ (1979) Information retrieval, 2nd edn. Butterworth-Heinemann, USA

Vapnik V (1998) The support vector method of function estimation. In: Suykens JAK, Vandewalle J (eds) Nonlinear modeling. Springer, Boston, MA, pp 55–85. https://doi.org/10.1007/978-1-4615-5703-6_3

Venkatesh B, Anuradha J (2019) A review of feature selection and its methods. Cybern Inf Technol 19:3–26

Vetschera R (2016) Concessions dynamics in electronic negotiations: a cross-lagged regression analysis. Group Decis Negot 25:245–265

Vetschera R, Koeszegi ST, Schoop M (2011) Electronic negotiation systems. In: Cochran JJ (eds) Wiley encyclopedia of operations research and management science, pp 1–8

Vijayarani S, Ilamathi MJ, Nithya M (2015) Preprocessing techniques for text mining—an overview. Int J Comput Sci Commun Netw 5:7–16

Wall ME, Rechtsteiner A, Rocha LM (2003) Singular value decomposition and principal component analysis. In: Berrar DP, Dubitzky W, Granzow M (eds) A practical approach to microarray data analysis. Springer, Boston, pp 91–109. https://doi.org/10.1007/0-306-47815-3_5

Weingart LR, Olekalns M (2004) Communication processes in negotiation: frequencies, sequences and phases. In: Brett J, Gelfand M (eds) The handbook of negotiation and culture, pp 143–157

Weingart L, Smith P, Olekalns M (2004) Quantitative coding of negotiation behavior. Int Negot 9:441–456. https://doi.org/10.1163/1571806053498805

Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemom Intell Lab Syst 2:37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Xie P, Xing EP (2013) Integrating document clustering and topic modeling. arXiv:1309.6874.

Xu D, Tian Y (2015) A comprehensive survey of clustering algorithms. Annu Data Sci 2:165–193. https://doi.org/10.1007/s40745-015-0040-1

Yan J, Zhang B, Liu N, Yan S, Cheng Q, Fan W, Chen Z (2006) Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing. IEEE Trans Knowl Data Eng 18:320–333. https://doi.org/10.1109/TKDE.2006.45

Yim O, Ramdeen KT (2015) Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. Quant Methods Psychol 11:8–21. https://doi.org/10.20982/tqmp.11.1.p008

Yuan G, Sun P, Zhao J, Li D, Wang C (2017) A review of moving object trajectory clustering algorithms. Artif Intell Rev 47:123–144. https://doi.org/10.1007/s10462-016-9477-7

Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed J (2020) A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. J Appl Sci Technol Trends 1:56–70

Zerhari B, Lahcen AA, Mouline S (2015) Big data clustering: algorithms and challenges. In: Proceedings of the international conference on big data, cloud and applications (BDCA'15)

Zhang W, Yoshida T, Tang X (2008) Text classification based on multi-word with support vector machine. Knowl Based Syst 21:879–886. https://doi.org/10.1016/j.knosys.2008.03.044

Zhu Y, Ting KM, Carman MJ (2016) Density-ratio based clustering for discovering clusters with varying densities. Pattern Recogn 60:983–997