Received: August 10, 2020. Revised: October 12, 2020.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Similarity based K-Means Clustering Technique for Categorical Data in Data Mining Application

Pradeep Kumar[1]* **Anita Kanavalli[1]**

[1]*Department of Computer Science and Engineering, Ramaiah Institute of Technology, India*
* Corresponding author's Email: pradeepkumard@msrit.edu

Abstract: Clustering plays a major role in the data mining application, because it divides and groups the data effectively. In the pattern analysis, two major challenges occur in real-life applications that includes handling the categorical data and the availability of correctly labeled data. According to the characteristics of homogeneity, the clustering techniques are designed to group the unlabeled data. Some important issues such as high memory utilization, time consumption, overhead, computation complexity and less effective results are present in various existing algorithms of numerical data. Therefore, the research study implemented clustering techniques based on the similarity of categorical data. Simultaneously, the attributes of inter and intra-clusters' similarities are identified, and then the performance of proposed method is improved by integrating those similarities. The noises are also removed by performing the pre-processing techniques, so the similarity between noise-free elements are estimated. Once these similarities are identified, the insignificant attributes are removed and the relevant attributes are chosen from the pre-processed elements. The overhead is reduced by developing the Similarity-based K-means Clustering (SKC) approach for clustering the attributes that depends on divergence distance. The efficiency of SKC is tested in the experimental analysis by means of precision, f-measure, accuracy, error rate of clustering and recall. The results state that the developed study achieved 98.45% accuracy for the publicly available dataset when comparing with the existing techniques: variations of Particle Swarm Optimization (PSO) and semi-supervised clustering system.

Keywords: Categorical data, Clustering technique, Computation complexity, Divergence distance, Labeled data.

1. Introduction

In machine learning and data mining, clustering is considered as an important technique to divide the dataset into multiple clusters without using prior knowledge. The main aim of clustering is to group the datasets into clusters, so that the objects having high similarity in the same clusters are grouped together, where other objects have dissimilarity is grouped in other clusters [1]. The dataset has an inherent structure, which is often uncovered by using clustering techniques [2], for instance, similar objects are grouped together or to a centroid [3]. The objects are described by using their attribute values that are identified by the similarities and dissimilarities of those objects. Researchers implemented numerous clustering algorithms like grid-based, partitional, density-based, model-based and hierarchical clustering [4, 5]. In general, the clustering algorithms consists of different data types like categorical, numeric, ordinal and so on. The categorical data didn't have any natural order in attributes, hence it is not suitable to apply the distance measures straightforwardly on those categorical attributes [6]. Hence, researchers in data mining fields faces the challenging and difficult task for clustering categorical data [7].

Even though, several algorithms are proposed for clustering the categorical data, it still exhibits important drawbacks namely, i) No clustering algorithms are able to provide better results for various data types [8, 9]. ii) Processing with a vast number of data and high-dimensional data instances are considered as one of the important challenges. iii) Matching the different datasets by choosing a suitable dissimilarity measure is also a difficult task [10]. The

cost time is highly increase by processing a transactional dataset, where the quality of outcomes is affected by analyzing the numerous iterations in the process. The feature selection and dimensionality reduction techniques are developed to address this issues, where the main aim is to remove the noisy, redundant and irrelevant information by pre-processing the data [11, 12]. Recently, K-means and its variants are highly used for clustering large datasets because of their higher scalability and efficiency. The Euclidean distance is used in the cost function; hence these methods are not suitable for clustering the categorical data [13]. In order to deal with this issue, the study uses the divergence measure for categorical clustering instead of Euclidean distance, where numerical data are converted into categorical data using similarity calculation. Pre-processing, mining and validation of results are the major steps involved in clustering, where unwanted data are removed or missing values are filled in the pre-processing step. The existing techniques perform the clustering process along with outlier data, where proposed SKC remove the outliers by calculating the similarity between the data, before clusters the data. The useful information is mined from the input data by performing various tasks includes inter- and intra-attribute similarity calculation sequentially. Finally, the categorical data are clustered by using K-Means algorithm with less clustering error rate. The validation of the proposed method is carried out by using various UCI datasets against the existing techniques.

The rest of this research paper consists of: Section 2 reviews the existing techniques with its advantages and limitations for categorical data. The explanation of the proposed method is illustrated in Section 3. The validation of proposed SKC with existing techniques for various datasets by means of several metrics are presented in Section 4. Finally, the conclusion of this work with future development is described in Section 5.

2. Literature review

In this section, a discussion of various existing techniques is presented, which are used to cluster the categorical data. The advantages and its limitations of these existing techniques [14-18] are also illustrated.

Nguyen, and Kuo [14] clustered the categorical data by developing an Automatic Fuzzy Clustering with Non-Dominated Sorting PSO as AFC-NSPSO algorithm. The number of clusters and partition data instances were automatically determined by the developed study. The multiple objectives of fuzzy clustering were handled by NSPSO, where the final solution was selected from the Pareto front and various internal CVI indices namely DB, Dunn and Sil. The experiments were carried out on UCI datasets to validate the effectiveness of AFC-NSPSO against existing techniques. However, outliers were not considered being processed before the clustering.

Pang [15] partitioned a large-scale dataset into numerous independent sub-datasets by implementing a MapReduce-based hierarchical sub-space algorithm. A data-partitioning strategy was used to couple the PAPU with attribute-value weights, where similar data objects were mapped in the datasets. During global clustering phase, sub-clusters were iteratively merged by applying the hierarchical scheme with PAPA measures. The developed study achieved high clustering efficiency on real-world large-scale and synthetic datasets was proved by the experimental results. The data skewness problem was not considered in this study, which was the major limitation.

Sarkar [16] integrated the machine learning technique with semi-supervised clustering technique called credibilistic measure or CrKMd. The homogeneity was identified by using credibilistic measure and then the coincident clustering problems were avoided. In the second part, a supervised model was built by using the clustered data to classify the unlabeled or uncertain data. The experimental results stated that the developed method provided better performance than other existing techniques by handling those ambiguous or uncertain data. However, the time and complexity of this work was not considered and outliers are not removed.

Sajidha [17] proposed the modified K-means clustering algorithm by considering every attribute of datasets for selecting the initial seed. The datasets were clustered with the mixed attributes easily because the developed study was independent of user-defined parameters. A novel distance measure technique was proposed to handle the numerical data and then the values were assigned as zero or one. The modified K-means algorithm was used to handle those mixed attributes, where the experiment was conducted to test the efficiency of the developed method with existing approaches. But, the developed study provided poor performance because the selection of outliers was not considered.

Dutta [18] discovered an optimal value of K by designing an automatic clustering algorithm. The iterative hill-climbing algorithms namely Genetic Algorithm (GA) and K-Means were used to identify the local and global optimum solutions. The non-linearly separable clusters were grouped by implementing a Multi-Objective GA as (MOGA). The intra-cluster distance was minimized and inter-

cluster distance was maximized by MOGA, where it was specially proposed for handling the mixed types of features. The superiority of MOGA algorithm was verified by conducting the experiments on UCI datasets. However, it was trying to capture only the spherical shaped clusters and it was prioritizing to considering importance of features of clustering, i.e. it is considering all features equally.

3. Proposed methodology

The numerical data are processed by developing various clustering algorithms and k-means is considered as one of the most important clustering algorithms. But, a vast amount of categorical data is presented in several applications namely text mining, market based data analysis and protein sequence analysis. When compared with the numerical clustering data, researchers face challenging issues for clustering the categorical data due to absence of inherent distance measure between categorical objects [19, 20]. In this research work, the SKC is used to cluster the categorical data by using a divergence measure. Fig. 1 shows the description of proposed method to extract the information from a vast amount of data. Initially, the noises or unwanted attributes are removed from the dataset. The numerical data are achieved by converting the categorical data once the estimation of similarities between intra and inter-attributes are conducted.

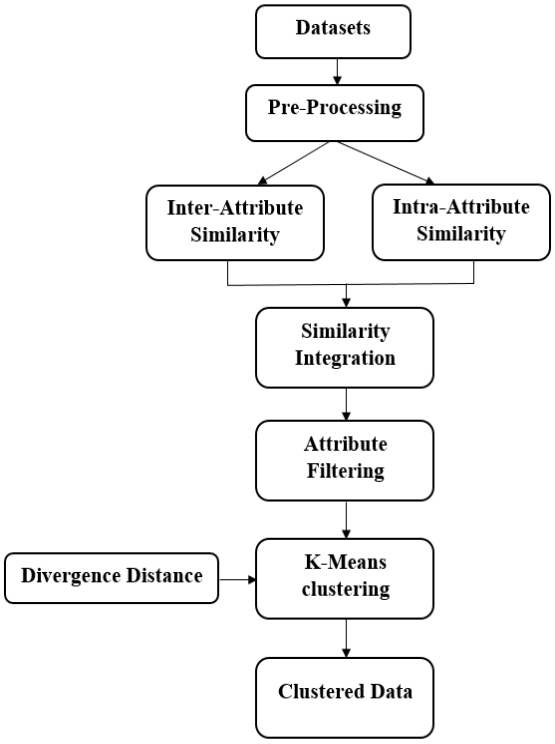

Figure. 1 Working Procedure of Proposed SKC Technique

Then, further processing is carried out by integrating the similarities of both attributes. Finally, according to the divergence distance, the similar attributes are grouped in the research study by using K-means clustering.

3.1 Pre-processing

In various data mining applications, pre-processing is the most important essential task for effective clustering. In order to improve the performance of the developed study, the noises or unwanted attributes are removed. For example, the adults contain various attributes namely gender, age, occupation, native country, education, salary, etc. that must be removed in this stage. Then, the distinct attributes are estimated for computing the similarities between noise-free attributes.

3.2 Computing the similarities

Nowadays, researchers considered similarity analysis is a challenging and tedious process in several domains. There are four major categories presents in the similarity-based clustering, which are stated as follows:

- Distance/Similarities between value of the attributes,
- Distance between different types of clusters,
- Distance between data objects, and
- Distance between attributes.

Once, the pre-processing stage is finished, the research study performed the intra- and inter-similarity analysis between those noise-free data. The similarity between the same column attributes are identified in the intra-attribute similarity, where the greater similarities are revealed according to the approximation of equal frequencies. When the frequency is high, then the distance between two closer values are estimated. In addition, the distinct levels of attribute value significance are indicated by various frequencies. The distance between rows are determined by selecting the inter-attributes. The couplings between the attribute values are not considered during process of calculating the attribute values. The attribute values are similar, only if they occur with the same relative frequency. Once the inter- and intra-similarity attributes calculation are finished, then the single attribute similarity is achieved by integrating those values. Finally, the filtering process will occur for further steps, where the similarity is reflected as discrepancies presented in the estimation of attribute value on the basis of frequency distributions. The combination of intra-

and inter- similarity measures are represented as the overall similarity. The maximum similarities are assigned by the attributes those are having equal frequency distributions. Therefore, the closeness is represented by similar frequencies and distinct levels are used to indicate the dissimilar frequencies.

3.3 Similar data calculation

The data can be clustered using k-means algorithm, but some outliers will present in these normalized data. The initial step is to calculate the similarity between the data by using divergence distance which is also used to remove the outliers.

3.3.1. Distance/similarity metrics

The quantitative degree of how far-off the two objects is defined by distance from the perspective of numeric and scientific applications. When the objects are similar, then the similarity value of those objects are higher. According to the distance between two objects of clusters, the cluster assignment of each data object is accomplished. Finally, the objects are assigned to the cluster, when it is nearest to the similar objects and dissimilar to the other objects of other clusters. In this proposed method, instead of Euclidean distance, the similarities of clusters are calculated by using divergence. Eq. (1) shows the distance measure of divergence as:

$$d_{Divergence} = 2\sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{(P_i + Q_i)^2} \quad (1)$$

Where, $P_i \; and \; Q_i$ are the two categorical data objects. By using the above equation, the similarity between two clusters are obtained. Then, these similar data can be clustered by using K-means algorithm which is described below.

3.3.2. K-means algorithm

According to the size and number of the data, the optimum number of desired clusters are identified by using k-means algorithm. Consider, the total number of data nodes is depicted as N that is uniformly distributed in a $M \times M$ square region. The mathematical expression Eq. (2) defines the optimum number of clusters k_{opt} that is defined as follows

$$k_{opt} = \frac{\sqrt{N}}{\sqrt{2\pi}} \sqrt{\frac{\varepsilon_{fs} M}{\varepsilon_{mp} {d_{x_i x_j}}^2}} \quad (2)$$

Where, the distance is illustrated as $d_{x_i.x_j}$ for the data $x_i \; and \; x_j$, the parameter for the free space model is described as ε_{f_s} and the parameter for multipath model is denoted as ε_{mp}. A given set of input data are classified into k number of disjoint clusters by using the k-means clustering algorithm and Eq. (2) is used to predefine the value of k. There are two phases presented in the basic k-means algorithm, which is described below:

Algorithm 1: The K-means clustering algorithm

Input: Initialize the set of n data items as D

Initialize the number of desired clusters is equal to k

Output: k clusters' set.

Steps:

1: Choose the initial centroids as k data-items randomly from D;
2: Repeat the process;
3: Each data item is allocated to its nearest centroid;
4: New mean is identified for every cluster;
5: Stop the process, once the convergence criteria are met.

From the above algorithm, the data can be clustered without outliers in an effective way. In the next section, the effectiveness of SKC can be validated by using various experiment analysis on UCI dataset as described.

4. Results and discussion

In this section, the validation of the proposed SKC method against existing techniques are illustrated with several parameter metrics. The experiments are carried out on publicly available datasets, which are explained in the below section.

4.1 Dataset description

The UCI machine learning datasets used in this work can be downloaded by referring to the following links that are described in Table 1. In addition, Table 2 describes the total number of attributes, number of instances and missing value of each dataset.

Table 1. Dataset with its link

Dataset	Links to download
Adult	https://archive.ics.uci.edu/ml/datasets/adult
Connect-4	https://archive.ics.uci.edu/ml/datasets/Connect-4
Chess	https://archive.ics.uci.edu/ml/datasets/Chess+(King-Rook+vs.+King)
Mushroom	https://archive.ics.uci.edu/ml/datasets/mushroom

Table 2. Datasets description

Dataset	No. of Attributes	No. of Instances	Missing Values
Adult	14	48,842	Yes
Connect-4	42	67,557	No
Chess	6	28,056	No
Mushroom	22	8124	Yes

The next section describes the validation metrics of the proposed method and then, the experimental settings are discussed.

4.2 Experimental setup and parameter metrics

The computer with 2.2 GHz of Intel Core i5, RAM of 8GB was used to implement the SKC method by using the programming language of Python 3.7.3 version. The performance of SKC method is validated by conducting several experiments on UCI dataset using various metrics namely clustering error, accuracy, F-measure, precision and recall.

The proportion of positive samples are classified as positive by using sensitivity rate i.e. true positive rate. In contrast with this, the negative samples are correctly classified as negative by using specificity measure i.e. true negative rate. Accuracy can be calculated by using Eq. (3), and Eq. (4) to evaluate the single combined metric, which is defined as F-measure. Among the number of labeled positive class samples, precision is used to identify the number of accurately labeled samples, which is shown in Eq. (5). On the contrary, according to the positive class, recall is used to predict the number of accurate positive class labeled samples, which can be divided by the total number of samples. The mathematical expression for recall is given in Eq. (6). Due to the process of clustering, the errors may occur and those errors are identified by the rate called clustering error. The mathematical equation for clustering is given in Eq. (7),

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

$$F - Measure = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (4)$$

$$Precision = \frac{TN}{TN+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

Where, TP is true positive, TN is true negative, FP is false positive and FN is false negative.

$$Clustering\ Error = \frac{\sum_{i=1}^{N}\delta(a_i, map(m_i))}{N} \quad (7)$$

Where, number of instances are represented as N in the dataset, a_i is used to describe the provided label and m_i illustrates the mapping function, which are used to map the obtained cluster label. The clustering error is computed as $e = 1 - Error$.

5. Performance analysis of proposed method by means of accuracy and clustering error

In this section, the validation of SKC method is analyzed against various existing techniques such as AFC-NSPSO [14], CrKMd [16] and other popular techniques like Support Vector Machine (SVM) and Naive Bayes (NB). The existing AFC-NSPSO and CrKMd conducted the experiments only on mushroom dataset. Therefore, to validate the SKC method on various datasets, this research work implements the AFC-NSPSO and CrKMd for other datasets and experiments are conducted. Table 3 shows the experimental results of SKC method and graphical representation is given in Fig. 2.

These experimental results clearly state that the SKC method achieved higher accuracy for four datasets when compared with existing techniques. While comparing with connect-4 dataset, all other datasets provided better results for the SKC method. For instance, the SKC method achieved nearly 98.5% accuracy for adult, chess and mushroom dataset. However, the existing techniques namely SVM and AFC-NSPSO achieved nearly 96% accuracy for the same three datasets.

Table 3. Comparative analysis of proposed SKC method

Methods	Accuracy (%)			
	Adult	Chess	Connect-4	Mushroom
SVM	96.40	96.90	77.46	96.90
NB	97.37	95.64	78.48	96.24
AFC-NSPSO	96.82	95.57	79.47	96.45
CrKMd	93.56	96.24	76.26	95.57
Proposed SKC Method	98.45	98.84	81.47	98.15

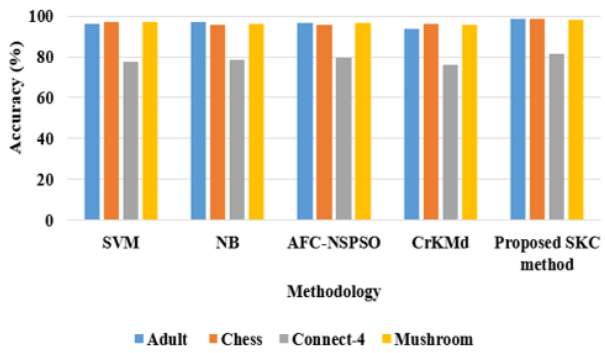

Figure. 2 Performance of SKC method in terms of accuracy

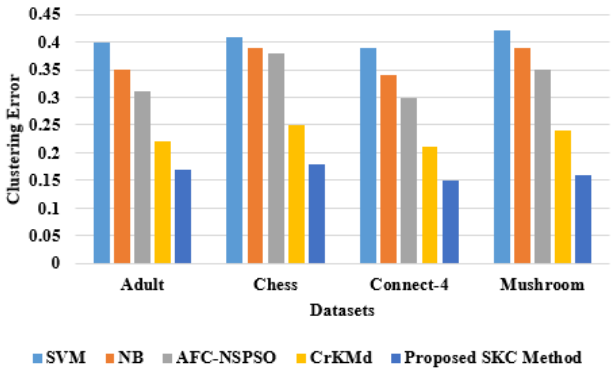

Figure. 3 Clustering Error Values of SKC Method

Table 4. Clustering error for proposed SKC method

Methods	Clustering Error			
	Adult	Chess	Connect-4	Mushroom
SVM	0.40	0.41	0.39	0.42
NB	0.35	0.39	0.34	0.39
AFC-NSPSO	0.31	0.38	0.30	0.35
CrKMd	0.22	0.25	0.21	0.24
Proposed SKC Method	0.17	0.18	0.15	0.16

For connect dataset, the existing techniques namely SVM, NB, AFC-NSPSO and CrKMd achieved only 76% to 79% of accuracy, where proposed SKC achieved 81.47% of accuracy. The existing AFC-NSPSO [14] and CrKMd [16] didn't consider the removal of outliers before clustering process, where SKC removed the outliers that leads high performance on accuracy. This is due to the distance measures used in the SKC method for clustering the data. Table 4 describes the clustering error of the SKC method against existing techniques such as NB, SVM, AFC-NSPSO and CrKMd. The graphical representation is shown in Fig. 3.

While clustering the data, the error may happen and this error can lead to poor performance in the analysis. From Table 4, it is clearly shows that the SKC method obtained less error rate in clustering when compared with other existing techniques namely SVM, NB, CrKMd and AFC-NSPSO for four datasets. For instance, the AFC-NSPSO obtained 0.30 error rate, NB achieved 0.34 error rate, SVM achieved 0.39 error rate and CrKMd obtained 0.21 error rate for the Connect-4 dataset. But, SKC method achieved only 0.15 error rate for the same dataset. In addition, the NB achieved nearly 0.34 to 0.39 error rate on all four datasets, where proposed SKC achieved nearly 0.15 to 0.18 error rate on all four dataset. This is due to the divergence distance used in this research work. In the next sub-section, the performance of SKC method is validated by using precision and recall metrics.

6. Analysis of proposed method in terms of precision and recall

In this section, parameters like precision and recall for SKC method are compared with existing techniques such as SVM, NB, AFC-NSPSO and CrKMd. The experimental results are tabulated in Table 5, in which the best values are making it as bold. Fig. 4 and 5 show the graphical representation of recall and precision of the SKC method with several existing techniques.

From the Table 5, it is clearly stated that the performance of SKC method achieved higher recall values for four datasets, when compared with existing techniques. However, all the techniques achieved less recall values on Adult and Connect-4 datasets, where SVM achieved nearly 70% to 79% of recall value, NB achieved nearly 75% to 83% of recall value, but SKC achieved 98.57% and 95.56% of recall on Adult and Connect-4 datasets. The reason is that the clustering process are effectively handled by k-means algorithm with distance metrics. When compared with AFC-NSPSO, SKC method improved 5% recall values for adult, 3% recall for chess, 1% recall for mushroom and 10% recall for connect-4 datasets. However, the SKC method achieved less performance in recall values for connect-4 datasets among all other datasets.

Table 5. Performance of SKC method in terms of precision and recall

Methods	Recall (%)				Precision (%)			
	Adult	Chess	Connect-4	Mushroom	Adult	Chess	Connect-4	Mushroom
SVM	78.61	96.86	69.07	96.86	82.58	96.89	89.62	96.86
NB	83.40	96.24	75.59	97.42	94.90	96.59	88.80	96.59
AFC-NSPSO	93.43	96.62	79.84	97.62	90.37	95.45	88.61	95.45
CrKMd	96.51	98.50	81.49	96.14	84.16	98.25	78.46	95.47
Proposed SKC Method	98.57	99.39	95.56	98.47	97.64	99.74	96.39	98.50

Table 6. Performance of F-Measure

Methods	F-Measure (%)			
	Adult	Chess	Connect-4	Mushroom
SVM	85.16	84.36	87.18	83.49
NB	86.48	86.07	88.97	90.24
AFC-NSPSO	87.05	92.94	89.83	92.84
CrKMd	92.52	96.15	91.15	94.18
Proposed SKC Method	95.47	99.07	93.62	97.21

The SKC method achieved higher precision values for all the four datasets, when compared with AFC-NSPSO and CrKMd, which is shown in Fig. 5. SKC method achieved approximately 97% to 99% precision for all the four datasets. The existing techniques achieved nearly 95% precision for mushroom dataset, where the SKC method achieved 98.50% precision for the same dataset.

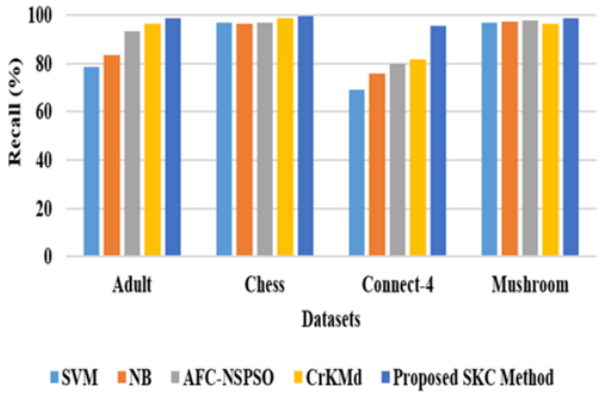

Figure. 4 Analysis of Recall for Proposed Method

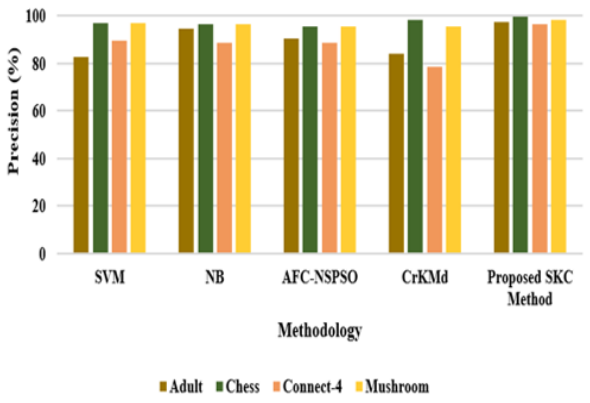

Figure. 5 Analysis of SKC Method in terms of Precision

The reason for achieving high precision value is that the outliers are effectively removed before clustering process using similarity distance in SKC, where outliers are not removed by the existing techniques. Finally, the f-measure of SKC method is validated in the next sub-section.

7. Analysis of proposed method by means of F-measure

The experiments were conducted on all datasets to validate the performance of the SKC method in terms of F-measure, which is shown in Table 6. The graphical representation for F-Measure of SKC method is compared with SVM, NB, AFC-NSPSO and CrKMd is described in Fig. 6. From the experimental analysis for f-measure, the results showed that SKC method achieved high performance than the existing techniques. For adult dataset, the existing techniques achieved 85.16%, 86.48%, 87.05% and 92.52% f-measure, but the SKC method achieved 95.47% f-measure.

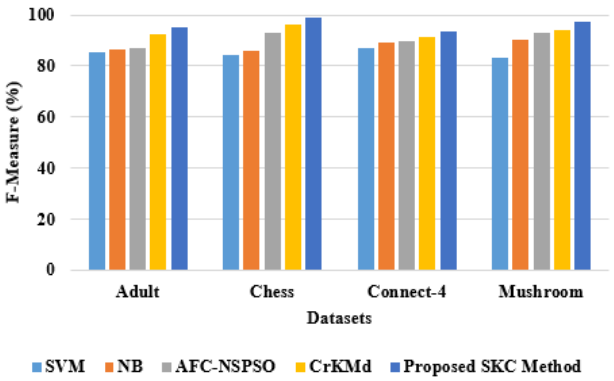

Figure. 6 Comparative Analysis of Proposed Method in terms of f-measure

For connect-4 dataset, the existing techniques achieved nearly 87% to 90% f-measure, where the proposed SKC achieved 93.62% of f-measure. The SVM is not suitable for large value dataset and NB assumes all the input attributes are mutually independent that leads poor performance on all the four datasets than AFC-NSPSO and CrKMd. The SKC method improved 5% f-measure than the AFC-NSPSO for mushroom dataset. Therefore, when compared with all the existing techniques for different datasets, SKC method achieved better performance in terms of accuracy, clustering error, precision, recall and f-measure.

8. Conclusion

The SKC techniques are implemented to process the large-scale datasets, where the main aim of the proposed method is to minimize the error rate of clustering during similarity computation. In the pre-processing step, the unwanted attributes from the adult database are removed. Once the noises are removed, the similarity values (i.e. inter and intra-similarity) between the attributes are identified. Then, k-means clustering technique is used to cluster the filtered data according to its divergence and similarity distance. The key benefits of the developed study are to minimize the time consumption, low computation complexity and high efficiency. The existing techniques, namely CrKMd and AFC-NSPSO are used to test the efficiency of proposed SKC technique. When compared to those techniques, the SKC technique presented better performance such as 98.84% accuracy, 99.39% recall, 99.74% precision, 99.07% f-measure and 0.18% clustering error rate for chess dataset. Moreover, the superiority of the proposed SKC technique is proved by conducting the experiments on various datasets. In future work, this study will develop effective optimization algorithms to select the optimized data.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1st author. The supervision and project administration, have been done by 2nd author.

References

[1] H. Jia and Y. M. Cheung, "Subspace clustering of categorical and numerical data with an unknown number of clusters", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 29, No. 8, pp. 3308-3325, 2017.

[2] T. H. T. Nguyen, D. T. Dinh, S. Sriboonchitta and V.N. Huynh, "A method for k-means-like clustering of categorical data", *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-11, 2019.

[3] N. Pang, J. Zhang, C. Zhang, X. Qin and J. Cai, "PUMA: Parallel subspace clustering of categorical data using multi-attribute weights", *Expert Systems with Applications*, Vol. 126, pp. 233-245, 2019.

[4] S. Chakraborty and S. Das, "k-Means clustering with a new divergence-based distance metric: Convergence and performance analysis", *Pattern Recognition Letters*, Vol. 100, pp. 67-73, 2017.

[5] W. Wei, J. Liang, X. Guo, P. Song and Y. Sun, "Hierarchical division clustering framework for categorical data", *Neurocomputing*, Vol. 341, pp. 118-134, 2019.

[6] I. Saha, J. Prasad Sarkar and U. Maulik. "Integrated Rough Fuzzy Clustering for Categorical data Analysis", *Fuzzy Sets and Systems*, Vol. 361, pp. 1-32, 2019.

[7] T. Nguyen, T. P. Quyen and R. J. Kuo, "Partition-and-merge based fuzzy genetic clustering algorithm for categorical data", *Applied Soft Computing*, Vol. 75, pp. 254-264, 2019.

[8] R. S. Sangam and H. Om, "K-modestream algorithm for clustering categorical data streams", *CSI Transactions on ICT*, Vol. 5, No. 3, pp. 295-303, 2017.

[9] R. S. Sangam and H. Om, "An equi-biased k-prototypes algorithm for clustering mixed-type data", *Sadhana*, Vol. 43, No. 3, pp. 37, 2018.

[10] L. Yuan, W. Wang and L. Chen, "Two-stage pruning method for gram-based categorical sequence clustering", *International Journal of Machine Learning and Cybernetics*, Vol. 10, No. 4, pp. 631-640, 2019.

[11] S. B. Salem, S. Naouali, and Z. Chtourou, "A fast and effective partitional clustering algorithm for large categorical datasets using a k-means based approach", *Computers & Electrical Engineering*, Vol. 68, pp. 463-483, 2018.

[12] N. Yuvaraj and C. Suresh Ghana Dhas, "High-performance link-based cluster ensemble approach for categorical data clustering", *The Journal of Supercomputing*, Vol. 76, No. 6, pp. 4556-4579, 2020.

[13] S. Zhu and L. Xu, "Many-objective fuzzy centroids clustering algorithm for categorical data", *Expert Systems with Applications*, Vol. 96, pp. 230-248, 2018.
[14] T. P. Q. Nguyen and R. J. Kuo, "Automatic Fuzzy Clustering Using Non-Dominated Sorting Particle Swarm Optimization Algorithm for Categorical Data", *IEEE Access*, Vol. 7, pp. 99721-99734, 2019.
[15] N. Pang, J. Zhang, C. Zhang and X. Qin, "Parallel Hierarchical Subspace Clustering of Categorical Data", *IEEE Transactions on Computers*, Vol. 68, No. 4, pp. 542-555, 2018.
[16] J. P. Sarkar, I. Saha, S. Chakraborty and U. Maulik, "Machine learning integrated credibilistic semi supervised clustering for categorical data", *Applied Soft Computing*, Vol. 86, pp. 105871, 2019.
[17] S. A. Sajidha, K. Desikan and S. P. Chodnekar. "Initial Seed Selection for Mixed Data Using Modified K-means Clustering Algorithm", *Arabian Journal for Science and Engineering*, Vol. 45, No. 4, pp. 1-19, 2019.
[18] D. Dutta, J. Sil and P. Dutta, "Automatic Clustering by Multi-Objective Genetic Algorithm with Numeric and Categorical Features", *Expert Systems with Applications*, Vol. 137, pp. 357-379, 2019.
[19] X. Zhao, J. Liang and C. Dang, "Clustering ensemble selection for categorical data based on internal validity indices", *Pattern Recognition*, Vol. 69, pp. 150-168, 2017.
[20] L. Bai and J. Liang, "Cluster validity functions for categorical data: a solution-space perspective", *Data mining and knowledge discovery*, Vol. 29, No. 6, pp. 1560-1597, 2015.