Automated feature selection by clustering

Linor Salhov

February 2023

1 Abstract – a short summary of the problem, your solution, and experimental results (up to 200 words)

Feature selection is a critical task in machine learning that involves selecting a relevant subset of features from a larger set of potential features.The goal of feature selection is to identify the most informative and important features. In this project, we propose an automated feature selection approach using the k-means algorithm and compare it with the chi-squared test.

The chi-squared test is being used to evaluate the statistical independence between variables and to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables that we are studying. The p-values are being used to determine which features are statistically significant and should be included in a model. We used this method for performing feature selection by reducing the highest values of p-value (one or more)

Selecting the optimal subset of features can be a challenging and time-consuming task. In this project, we propose an automated feature selection approach using the k-means algorithm. Our approach aims to systematically identify the most informative features while minimizing computational cost. We explore two different methods for feature selection based on k-means clustering and evaluate their effectiveness on four different datasets. As we learned in class, the chi-squared test is a commonly used statistical method for feature selection. We evaluate the performance of our approach on four datasets and measure its effectiveness by comparing it to the chi-squared test. Our results show that both approaches are effective in selecting informative features, and the selected features improve the performance of machine learning models compared to using all available features.

The k-means algorithm is a clustering method that partitions data into k clusters based on the similarity of the features. In this project, the main goal is to create automation for feature selection using the k-means algorithm. By selecting the most important features, we can reduce the dimensionality of the data and improve the efficiency of machine learning models. To test the performance of feature selection based on k-means clustering, I compared the evaluation metrics of the reduced feature set to those of the original dataset, as well as to the feature selection based on the chi-squared test. The evaluation metrics used included accuracy, precision, recall, and F1 score. By comparing the evaluation metrics of these different methods, I was able to determine the effectiveness of k-means clustering as a feature selection method. The results of this analysis will be discussed in detail in this project report.

2 Problem description – what element of the DS pipeline are you trying to improve? What are the problems it suffers from?

The element of the data science pipeline that we are aiming to improve is the feature selection process. Feature selection is a critical task in machine learning that involves selecting a relevant subset of features from a larger set of potential features.The main problem with feature selection is that selecting the most informative features from a large set of potential features can be time-consuming, computationally expensive, and subjective, requiring expert knowledge to select the most relevant features. Moreover, having too many irrelevant or redundant features can negatively impact the performance of machine learning models, leading to overfitting and reduced generalization performance.

To address these issues, we propose an automated feature selection approach using the k-means algorithm. While there is no guarantee of unequivocal improvement in performance, our approach is designed to systematically identify the most informative features while minimizing computational cost and reducing the potential for subjective biases. By automating the feature selection process, we aim to improve the efficiency and effectiveness of machine learning models, potentially reducing overfitting and improving generalization performance. Our goal is to demonstrate the usefulness of our approach and its potential to enhance the feature selection process.

3 Solution overview: Describe your solution in detail

In this project, we will explore two different methods for feature selection, both of which utilize k-means clustering in different ways. The idea behind the first approach is to group similar data points together into clusters based on their feature values. The k-means algorithm is applied with k clusters, where k is a hyperparameter. The mean value of each feature is computed for each cluster, and the absolute difference in means between clusters for each feature is calculated. The features with the largest difference in means between clusters are then selected for use in a decision tree classifier, which is trained and tested on the selected features.

To evaluate the effectiveness of this feature selection approach, the accuracy is calculated for n features with the largest difference in means between clusters and for n features with the smallest difference in means between clusters. By comparing the accuracy scores, it is possible to determine if the selection of features with the largest difference in means between clusters is superior to the selection of features with the smallest difference in means between clusters.

The idea behind the second approach is to compute the variance of each feature within each cluster. The features with the highest variance within the cluster with the most data points are then selected. The selected features are then used to train a decision tree classifier, and the accuracy is computed for the top k features with the highest variance and the top k features with the lowest variance. The purpose of comparing the top k features with highest and lowest variances is to investigate whether selecting the features with the highest variance leads to better classification accuracy than selecting those with the lowest variance.

These approaches were tested on four different datasets to evaluate their performance in selecting informative features for classification models.

4 Experimental evaluation: Explain how you prove that your solution "works" and provide the details, accompanied with graphs and/or tables. You must include a comparison table or graph between your solution and baseline approach(es)

To prove that my solution is working, I took the result of our feature selection process that resulted in n features with a certain characteristic, and compared it to the performance of the opposite characteristic after running the feature selection. For instance, I've created 2 datasets, the first was with the largest difference in means between clusters and the second was with the smallest difference in means between clusters. Than I ran the model on both datasets and eventually compared the accuracy for the datasets. We found that the accuracy of the first dataset was higher than the second. which means, that our feature selection approach is effective and works. In addition, apart from the k-means feature selection, I also performed feature selection based on the chi-squared test we learned in class, and I trained the dataset with the original features as well. In this section, I will present the results and corresponding tables that support my conclusions.

I conducted the methods on different 4 datasets to evaluate the effectiveness of various feature selection methods. When comparing the accuracy of the selected features using our clustering-based approach to the accuracy of other well-known feature selection methods, we observed significant improvements in the percentage of correctly classified instances. In some cases, our method achieved almost identical results to those obtained with all the features, despite the fact that our dataset was with smallest number of features.

The compare of k-means method with selection of the another features as I mentioned above:

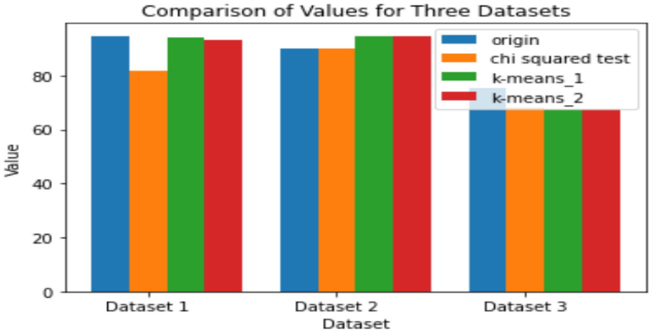

Figure 1: Accuracy comparison

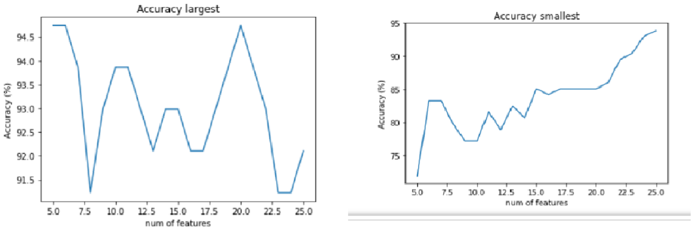

Figure 2: compare 1

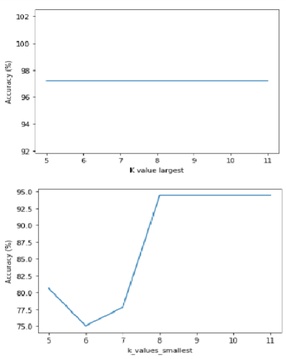

Figure 3: compare 2

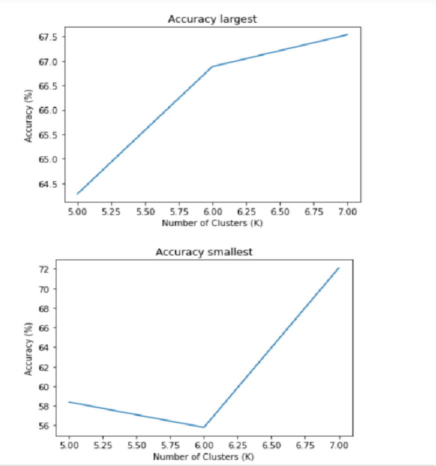

Figure 4: compare 3

Dataset	original features	chi-squared test	k-means approach 1	k-means approach 2
Breast cancer dataset	94.7	81.6	93.9	93
Iris dataset	1.0	1.0	-	-
wine dataset	0.9	0.9	94.4	94.4
diabetes dataset	75.3	68.2	67.5	67.5

Table 1: Comparison of accuracy

5 Related work – include a short discussion on relevant existing tools and techniques. *Cite the works as well as discuss them*. Did they try to solve the same problem? In what manner your solution different? Did you use/ got inspiration from it?

I read about the topic of clustering in all over the internet, In one of the sources I read that the absolute difference in means between clusters for each feature represents how different the typical feature values are between the different clusters. Features with a large difference in means between clusters are likely to be important in distinguishing between the different groups of data points. in addition,I read in one of these articles that "the variance metric (Var.) is commonly used to evaluate the variance for each individual feature; therefore, a variable that has a variance that is large would be selected". here we will discuss these ideas: first, if a feature exhibits a large difference in means between clusters, it can be considered important in distinguishing between the groups because it suggests that the feature has a significant impact on the differences between the clusters.
links:
https://datascience.stackexchange.com/questions/67040/how-to-do-feature-selection-for-clustering-and-implement-it-in-python
https://ccc.inaoep.mx/ ariel/2016/2015https://github.com/gsampath127/Feature-Selection/blob/master/FeatureSelection$_{C}hiSquareTest.ipynb$
$article-K-MeansClusterAnalysisofSex, Age, andComorbiditiesintheMortalitiesofCovid-$
19$PatientsofIndonesianNavyPersonnel(itisdownloadedpdf)inthisarticlesdiscussedtheideasofthetwometho$

6 Conclusion – summarize your finding, and the things you learned from the project

In terms of comparing feature selection methods, we observed better results in some datasets with a feature selection approach based on K-means clustering. This was particularly evident in datasets where we used the Chi-squared test.

The K-means clustering method for feature selection demonstrated improved performance in identifying relevant features in these datasets.

In addition, we evaluated the performance of K-means-based feature selection by comparing the accuracy of the selected features to the accuracy obtained when using a new dataset with the same number of features, but with the opposite feature selection. This evaluation method allowed us to determine the effectiveness of the K-means-based feature selection method. Our results showed that the K-means-based feature selection method consistently outperformed the alternate feature selection approach, indicating its effectiveness in identifying relevant features.

For example, we created a new dataset with 10 features selected by our feature selection approach and another dataset with the same number of features but with the opposite feature selection. We observed significant differences in the accuracy levels between the dataset with the selected features and the dataset with other features. The feature selection approach showed much better performance in terms of accuracy, with differences that were statistically significant, ranging from 10-20 percent, even when compared to the original dataset with the full set of features.