Classification of heterogeneous data based on data type impact on similarity

Najat Ali, Daniel Neagu, Paul Trundle

Artificial Intelligence Research (AIRe) Group
Faculty of Engineering and Informatics
University of Bradford, Bradford, UK
N.Ali50@Bradford.ac.uk, D.Neagu@Bradford.ac.uk, P.R.Trundle@Bradford.ac.uk

Abstract. Real-world datasets are increasingly heterogeneous, showing a mixture of numerical, categorical and other feature types. The main challenge for mining heterogeneous datasets is how to deal with heterogeneity present in the dataset records. Although some existing classifiers (such as decision trees) can handle heterogeneous data in specific circumstances, the performance of such models may be still improved, because heterogeneity involves specific adjustments to similarity measurements and calculations. Moreover, heterogeneous data is still treated inconsistently and in ad-hoc manner. In this paper, we study the problem of heterogeneous data classification: our purpose is to use heterogeneity as a positive feature of the data classification effort by using consistently the similarity between data objects. We address the heterogeneity issue by studying the impact of mixing data types in the calculation of data objects' similarity. To reach our goal, we propose an algorithm to divide the initial data records based on pairwise similarity for classification subtasks with the aim to increase the quality of the data subsets and apply specialized classifier models on them. The performance of the proposed approach is evaluated on 10 publicly available heterogeneous data sets. The results show that the models achieve better performance for heterogeneous datasets when using the proposed similarity process.

Keywords: heterogeneous datasets, similarity measures, two-dimensional similarity space, classification algorithms

1 Introduction

Data classification is an important topic in data mining. Plenty of classifiers have been proposed for classifying data objects according to some constraints and requirements [1]. In the real world, data is heterogeneous: a mixture of numerical and categorical features; classifying such data using existing methods may lead to possible misclassifications and open-ended issues, due to the nature of heterogeneous data. Practically heterogeneity is seen in the process of classification as a special type of contamination, making it difficult to build credible and consistent classification model(s). The main challenge for classifying heterogeneous datasets is how to deal with mixture of data types present in the dataset. We attempt to solve this issue by studying the impact of data similarity by their types on classifying instances from heterogeneous data sets.

The purpose of this paper is to utilize the influence of similarity measures on classification accuracy for heterogeneous data sets by generating a two-dimensional similarity space and classifying the data based on its similarity data values. Our motivation is to reduce the initial noisy data collection to more consistent subdomains that have all their data as similar as possible. Therefore, we first review the main notions of dissimilarity/similarity measures and present some of currently most known classification methods, and then we propose a new method to classify heterogeneous data set based on the newly introduced concept of the two-dimensional similarity space.

The rest of this paper is organized as follows: the next section provides the concepts, background and literature review relevant for the paper topic. Section 3 introduces the idea of the proposed similarity-based modeling. Section 4 reports experimental work and analysis of the results. Finally, Section 5 presents conclusions and future work.

2 Background

2.1 Distances and similarity measures

Many data mining algorithms use distance measures to determine and apply the similarity /dissimilarity (i.e. distance) between data objects. Similarity (and complementarily distance) functions are used to measure the degree to which data objects are comparably close (or not) to another [1].

Definition 1: Let A be a set of d-dimensional observations (e.g. data objects). A mapping $d : A \times A \rightarrow R$ is called a **distance metric** on A [2] if, for any $x, y, z \in A$, it satisfies requirements on:

1. $d(x,y) \geq 0$ $(non - negativity)$;
2. $d(x,y) = 0 \quad if \ x = y$ $(identity)$;
3. $d(x,y) = d(y,x)$ $(symmetry)$;
4. $d(x,z) \leq \mathrm{d}(x,y) + d(y,z)(triangle\ inequality)$.

Definition 2: Let A be a set of d-dimensional observations. A mapping $s : A \times A \rightarrow R$ is called a **similarity** on A if it satisfies the following properties [2]:

5. $0 \leq s(x,y) \leq 1$ $(non - negativity)$;
6. $s(x,y) = 1$ if $x = y$ $(identity)$;
7. $s(x,y) = s(y,x)$ $(symmetry)$.

A dissimilarity is generally a complementary mapping to the similarity definition. Plenty measures have been proposed for comparing data objects of same type in data mining applications. Some most popular distances for numerical data include Minkowski, Euclidean, Manhattan, and Chebyshev distances. The most common distances for categorical data types include Simple matching, Eskin, Tanimoto, Cosine and Goodall distances; more information about these distances can be found for example in [1] [3, 4]. For comparing objects described by a mixture of features using a specific distance or similarity measures the area is not that rich; a general similarity coefficient measure proposed by Gower in [5] is the most common measure for comparing such data [3]. Ottaway in [6] highlighted some of the problems involved. Because of the additional challenges representation, the similarity for heterogeneous data is more complicated; researchers in different data mining studies have used a combination approach for computing the distance by combining different distances for different data types.

In our study, we define heterogeneous data as a combination of a mixture of features, some are numerical, and some are categorical at least; there may be examples using other data types, but we did not consider them hereby. This paper tackles the classification problem of heterogeneous data as a mixture of numerical and categorical records with variations in either or both types. For the sake of simplicity, we apply Minkowski distance for comparing numerical features and simple matching distance for comparing categorical features; both of them satisfy distance *Definition 1* above. Minkowski and simple matching distances deal with the measurement of divergence between data objects, their similarity is calculated using relevant conversion methods.

2.2 Background: similarity in classification algorithms

The classification problem in data mining is a supervised machine learning task that approaches the recognition of a given set of entries by a label based on previously presented samples. Many different algorithms have been proposed for solving the classification problem based on a variety of techniques and concepts, for example most commonly used methods for data classification tasks include decision trees such as ID3[7], CART[8], C4.5[9], K-nearest neighbour (KNN) [10], Artificial Neural Networks (ANN) [11], Support Vector Machines (SVMs) [12], and Naïve Bayes[13].

In many different studies, researchers have used the above-mentioned methods for classifying data described by a mixture of numerical and categorical features by initially transforming data (pre-process step) before or during the classifier training steps; an example of these studies include [14, 15] and relevant examples are described below.

Some authors studied the problem of heterogeneous data classification by improving existing classifiers to handle heterogeneous data. In [16] Pereira et al. have proposed a new distance for heterogeneous data which is used with a KNN classifier. This distance, called Heterogeneous Centered Distance Measure (HCDM), is based on a combination of two techniques: the proposed method relies on dividing the data set into pure numerical and pure categorical features, then applies Nearest Neighbor Classifier CNND distance to numerical features and Value Difference Metric to categorical features, and the result of the two distances is assembled in one single distance to form the HCDM value.

In [17] Jin et al. proposed a novel method for heterogeneous data classification called Homogeneous data In Similar Size (HISS); their method is based on dividing heterogeneous data into a number of homogeneous partitions of similar sizes. Although their method showed a good performance for heterogeneous data classification, the authors did not consider the effects of homogeneous subsets on all relevant subspaces during training stage. Hsu et al in [18] studied a mixed data classification problem by proposing a method called Extended Naïve Bayes (ENB) for mixed data with numerical and categorical features. The method uses the original Naive Bayes algorithm for computing the probabilities of categorical features: numerical features are statistically adapted to discrete symbols taking into consideration both the average and variance of numeric values. In [19] Li et al. proposed a new technique for mining large data with mixed numerical and nominal features. The technique is based on supervised clustering to learn data patterns and use these patterns for classifying a given data set. For calculating the distance between clusters. The author have used two different methods; the first method was based on using specific distance measure for each type of features, and then combined them in one distance. The second method was based on converting nominal features into numeric features, and then numeric distance is used for all features. In [20] Sun et al. presented a soft computing technique called neuro-fuzzy based classification (NEF-CLASS) for heterogeneous medical data sets; the motivation at that time was based on the fact that most conventional classification techniques are able to handle homogeneous data sets but not heterogeneous ones. Their method has been tested on both pure numerical and mixed numerical and categorical datasets.

To summarise, the most commonly used approaches for handling data described by a mixture of numerical and categorical features use two approaches: (1) conversion methods of initial data components to a consistent standard data type for which relevant, specialized machine learning techniques are applied. For example, k-NN works naturally with numerical data, for heterogeneous data, the non-numerical data subset is converted into numerical data and sometimes calibrate/normalize or project that numerical data to reduce effects of disparate ranges. Alternatively, decision trees can be applied to heterogeneous data by converting numerical data into categorical data, and Naive Bayes are applied to learn discrete numeric attributes data converted into symbols. However, converting categorical features into numerical features (for example for SVM applications), may lead to loses of some useful information, a possible source of biased, or misclassification outcomes. (2) the hybrid ensemble development of classifiers by application of machine learning techniques to same data type component subsets, followed by a weighted average of all classifiers similar to computing the overall similarity value as a weighted average of same type data components. Each approach comes with added computational complexities and the need of data understanding and expertise to convert consistently either a priori or a posteriori the classification output.

3 Two-dimensional similarity space feature selection-based classification filter

In the proposed method, we intend to study the noise added by the numerical attributes and categorical attributes respectively, to the pairwise similarity of data records. This is approached by separating numerical features on one side, and categorical ones on the other side, and exploring when one becomes noisy for the other one, to leave just the case that they can still stay together when indeed full records are extremely similar.

Let $A = \{A_1, A_2, A_3, \ldots., A_N\}$ denote a set of d-dimensional objects of cardinality N, where each data object $A_i,\ i = 1,2,3, \ldots..N$, has d mixed features: d_1 numerical features $\{x_1, x_2, \ldots.x_{d_1}\}$, and d_2 categorical features $\{y_1, y_2, \ldots.y_{d_2}\}$, where $d = d_1 + d_2$ (for sake of presentation clarity the indexes of the above-named features are ordered).

For each feature type, one relevant distance mapping is applied, to create the two-dimensional similarity space 2DSS. Each point $\mathrm{Z_{ij}}$ in 2DSS is a pair of numerical and categorical similarity values $\mathrm{Z_{ij}} = (s_{N_{ij}}, s_{C_{ij}})$, where $0 \le s_{N_{ij}} \le 1$ and $0 \le s_{C_{ij}} \le 1$.

We define our similarity matrix (SM) as follows:

$$\mathrm{SM}=\begin{bmatrix} (s_{N_{11}}, s_{C_{11}}) & (s_{N_{12}}, s_{C_{12}}) & \cdots & (s_{N_{1n}}, s_{C_{1n}}) \\ (s_{N_{21}}, s_{C_{21}}) & (s_{N_{22}}, s_{C_{22}}) & \cdots & (s_{N_{2n}}, s_{C_{2n}}) \\ \cdot & \cdot & \cdots & \cdot \\ \vdots & \ddots & \vdots & \vdots \\ \cdot & \cdot & \cdots & \cdot \\ (s_{N_{n1}}, s_{C_{n1}}) & (s_{N_{n2}}, s_{C_{n2}}) & \cdots & (s_{N_{nn}}, s_{C_{nn}}) \end{bmatrix} \qquad (1)$$

SM is a symmetric matrix, the total number T of pairwise relevant points in the similarity space can be computed as $\mathrm{T} = \frac{\mathrm{N(N-1)}}{2}$.

The two-dimensional similarity space $2DSS$ is divided into four subspaces (see Figure 1). Subspace A contains all points $\mathrm{Z_{ij}} \in 2DSS$ with high similarity values for both numerical and categorical features $\mathrm{S_{NC}}$. Subspace B contains all $\mathrm{Z_{ij}} \in 2DSS$ that have a (relatively) high similarity value for numeric features, and low similarity values for categorical features $\mathrm{S_{N\bar{C}}}$. Subspace C contains all points $\mathrm{Z_{ij}} \in 2DSS$ with low similarity values for both numerical and categorical features$\mathrm{S_{\bar{N}\bar{C}}}$, and Subspace D contains all points $\mathrm{Z_{ij}} \in 2DSS$ that have low similarity values for numerical features, and high similarity values for categorical features $\mathrm{S_{\bar{N}C}}$. Figure 1 shows the division of the bi-dimensional similarity space in four relevant subspaces.

The proposed approach defines four directions for the original heterogeneous dataset to address the initial issues discussed in Section 2: data in subspace A is more homogeneous and requests a hybrid ensemble classifier or similar conversion methods that should learn data of high similarity; subspace C has noisiest samples that can be treated as outliers, subspaces B and D allow development of consistent, single-type machine learning models since data features of either numerical (subspace B) or categorical (subspace D) type are highly similar.

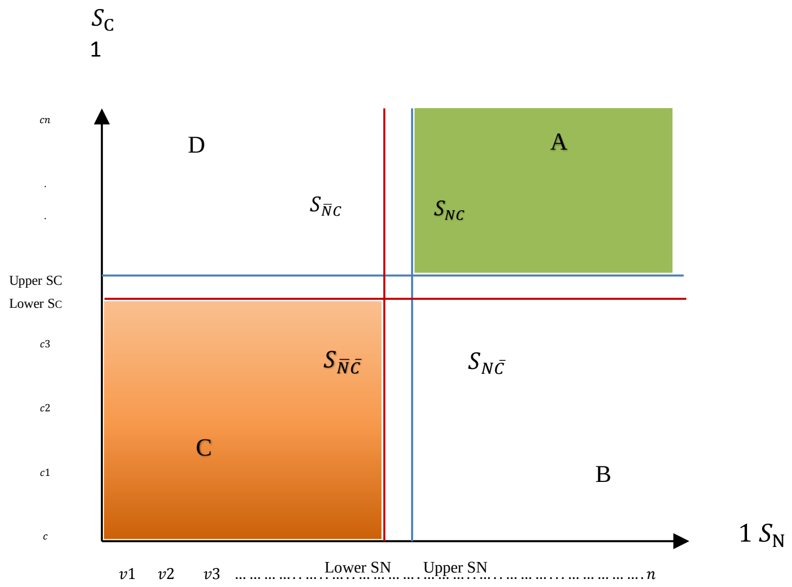

Figure 1. *The division of the two-dimensional similarity space 2DSS.*

The successful selection of the subspaces based on the proposed filter depends on the boundaries Lower_SN, Upper_SN that can be moved right or left across the continuous numerical similarity values domain (e.g. $v1, v2, \dots \dots .. n$) on the S_N axis, and the boundaries Lower_SC and Upper_SC that can be moved up and down across the discrete categorical similarity values $c1, c2, \dots \dots .. cn$ on the S_N axis as shown in Figure 1. The choice of these values can be a further optimization exercise. If these boundary values reach the maximum position then there will be just outliers, in the opposite case there will be no outlier cases. In the experiments reported in this paper these boundaries values are chosen by the rule of thumb.

Figure 2 shows the proposed methodology that allows filtering (extraction) of highly similar records without dependence of the classification outputs (classes or labels).

4 Experimental work

We performed our experiments on ten relevant heterogeneous datasets: three data sets from UCI Machine Learning Repository [21], and seven data sets from R packages datasets available in [22]. Each dataset contains different numbers of instances, attributes, and classes. A summary of properties of each data set is given in Table 1.

Table 1. Summary of data sets properties.

Dataset	observations	Numerical features	Categorical features
Student alcohol consumption	1044	9	22
Credit Approval	690	6	9
German Credit risk	1000	7	13
Structure of Demand for Medical Care	5574	9	5
treatment	2675	5	4
Visits to Physician Office	4405	10	8
Saratoga Houses	1728	10	5
Job train	2675	10	9
Labour training Evaluation1	15992	5	4
Wages and schooling	2944	10	16

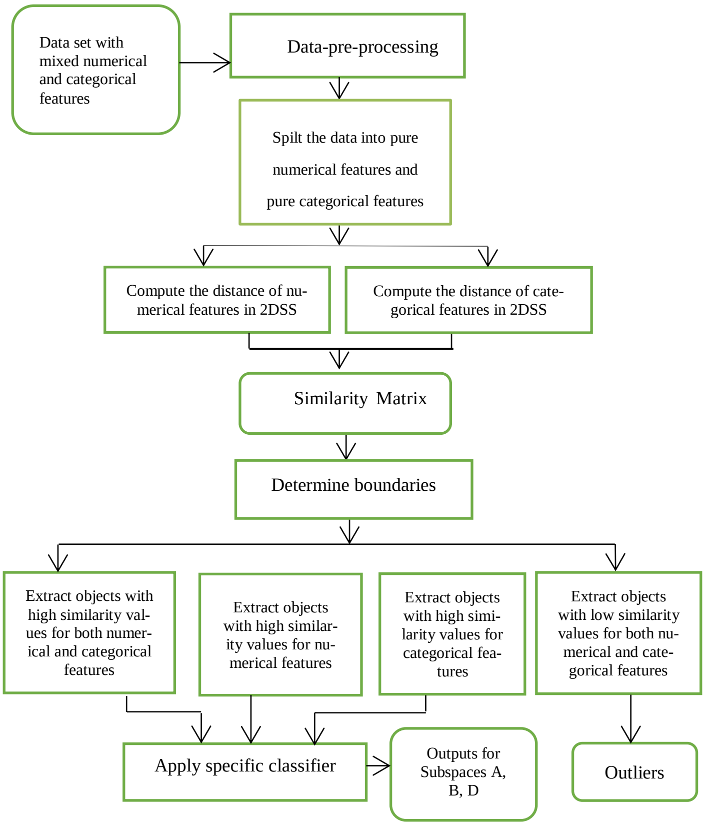
Data set with mixed numerical and categorical features
Data-pre-processing
Spilt the data into pure numerical features and pure categorical features
Compute the distance of numerical features in 2DSS
Compute the distance of categorical features in 2DSS
Similarity Matrix
Determine boundaries
Extract objects with high similarity values for both numerical and categorical features
Extract objects with high similarity values for numerical features
Extract objects with high similarity values for categorical features
Extract objects with low similarity values for both numerical and categorical features
Apply specific classifier
Outputs for Subspaces A, B, D
Outliers

Figure 2. *The proposed classification filter stages.*

All datasets are pre-processed before we ran the experiments, erroneous, inconsistent, and missing entries being removed. Data columns with more than 10% missing values are removed; the ordinal features are also removed from the data sets. Numeric features are normalized. K-fold cross validation method has been used for model. For validation, the original data is randomly partitioned into k equal size subsamples, where k =10.

Each data set has generally a limited number of categorical similarity values and a large number (practically any value in the numerical similarity domain0 of numerical similarity values, due to the intrinsic definition of these similarities. For example, Saratoga Houses data from table 1, has categorical similarity values (0, 0.2, 0.4, 0.6, 0.8, 1), and any numerical similarity value between 0 and 1s. We defined the categorical similarity boundary value where the performance of the model starts increasing.

4.1 Results

Classifiers' performances are compared based on their accuracy. Decision Tree C5.0 classification was first applied to all datasets. The results are shown in Figure 3:

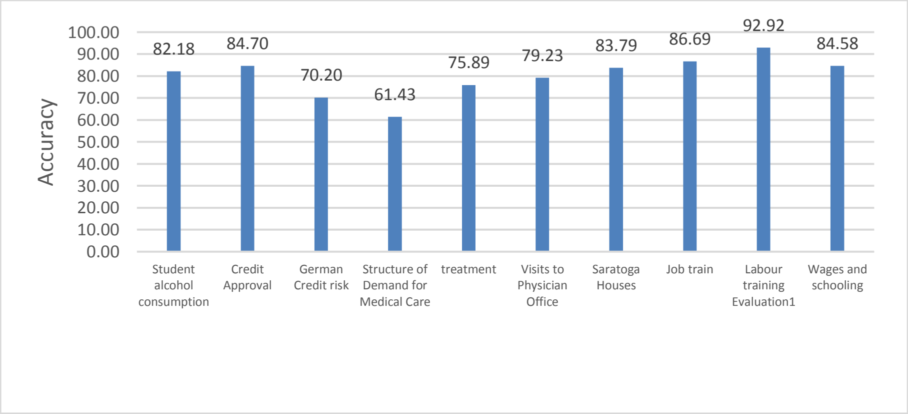

Figure 3. *Accuracy obtained by the classifier*

As mentioned earlier, for each data set, the similarity values of numerical and categorical features are represented as coordinate pairs in the 2DSS space. The performance of the filtering technique applied on the similarity space to define the four subspaces A, B, C and D for the next action (such as classification or outlier identification) depends on the boundaries selection. This is exemplified within the experiments with Decision Tree C5.0 algorithm applied to each data records extracted from each subspace. For pure numerical and pure categorical subspaces cases (i.e. subspaces B, and D, respectively) feature selection has been also applied to the filtered attributes. Data objects with low similarity values for both numerical and categorical values (subspace C) $S_{\overline{NC}}$ are considered outliers and separated from the main classification exercise.

Figures 4-8 show the results for the proposed method to the benchmark data sets. The models examined in the current experimental work perform well on the subsets of mixed numerical and categorical records with high similarity (subspace A) where data is homogeneous and similarity values S_{NC} are high. The improvement exceeded 4%, reaching a maximum of 14 % for the Student alcohol consumption dataset. In some cases though such as MedExp, Visits to Physician Office, and Labour training Evaluation1 datasets, the classifier performance increased just for the pure numerical features; also the performance of the classifiers increased just for pure categorical features in the case of Crx, German credit data set dataset. One of the reasons for such limited increase in subspace classifier performance is related to the chosen similarity distance.

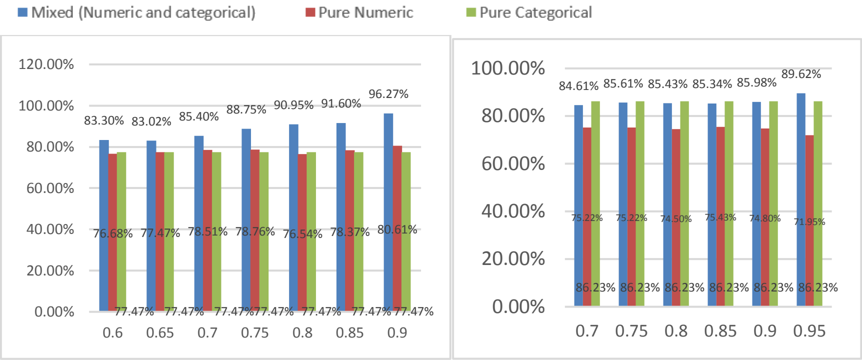

Figure 4 *Results for Student alcohol consumption and Crx datasets*

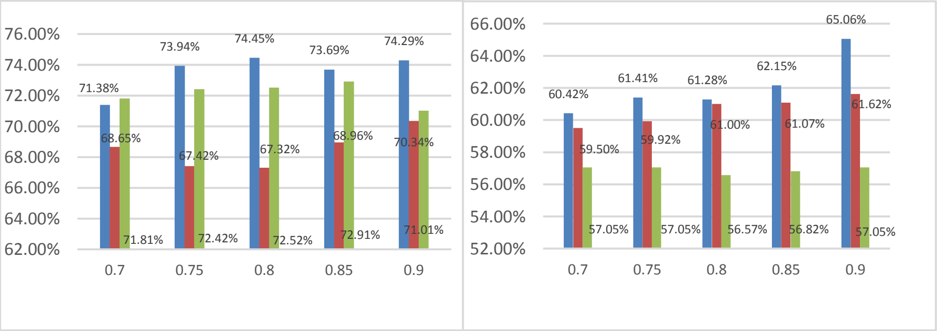

Figure 5 *Results for German credit card and MedExp datasets*

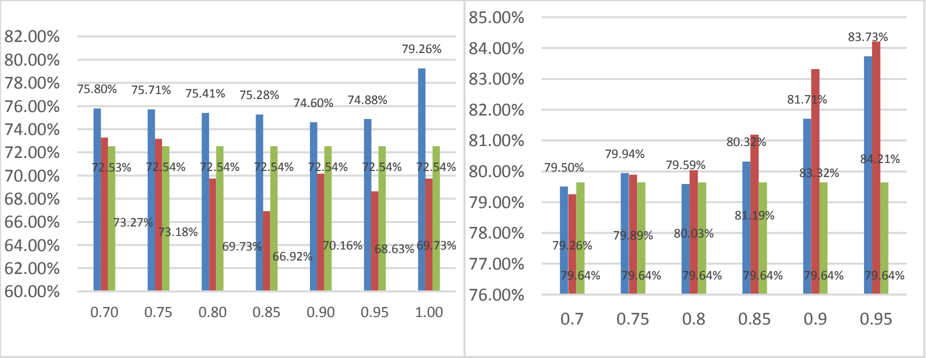

Figure 6 *Results for Treatment and OFP data sets*

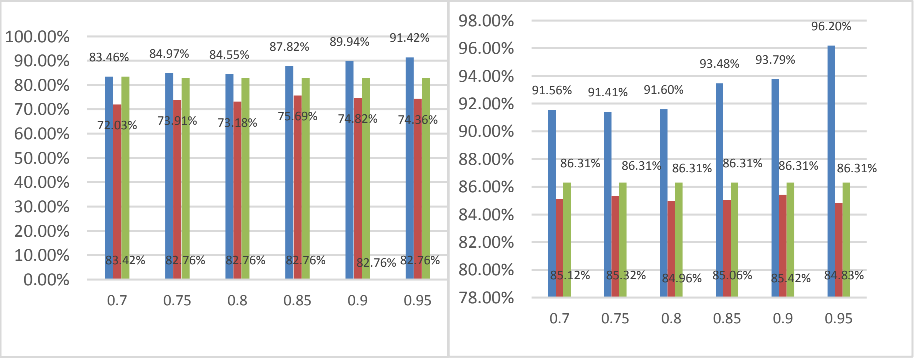

Figure 7. *Results for Saratoga Houses and Job train datasets*

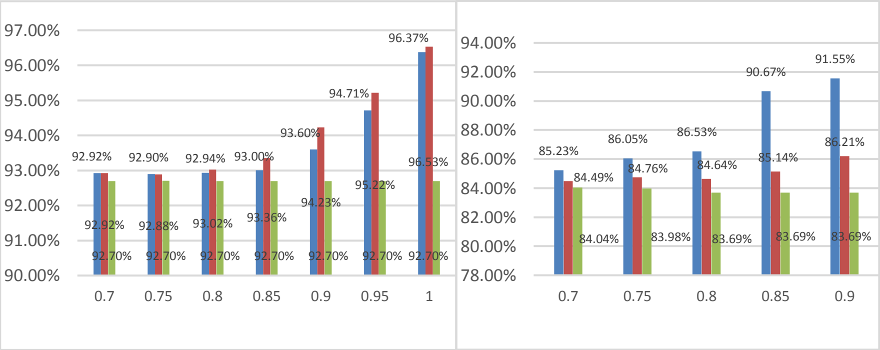

Figure 8 *Results for CPS1 and Schooling datasets*

5 Conclusions and further work

A new approach to filter records for classifying heterogeneous datasets based on the impact of data types on their similarity is proposed and evaluated. The similarity space is built in the experiments using a Minkowski distance for numeric features and simple matching for categorical features. The influence of similarity measures on the performance of the classifiers was investigated by identifying and removing the outliers (data objects with overall low similarity values) from the initial data set. In the experiments the influence of the similarity measures was investigated on classification accuracy.

It is important to point out that the proposed model may not handle efficiently small data sets because the subspaces become less relevant, and therefore we aim to investigate the performance and applicability of the proposed method for big heterogeneous data sets. There are also other wide areas for further work on this topic. We have tested currently the method on data samples with limited types of features (numerical and categorical) only. Future work may include other types of data features, such as ordinal, nominal, binary, and fuzzy, and extend the similarity space to multidimensional similarity space forms. We have used the most common distances (Minkowski and simple matching) for computing data objects similarities. More studies should be done to investigate the impact of the choice of similarities on the performance of the model. In addition, each subset may request optimisation of the specific classifier instead of applying just one classifier algorithm. The outliers can be also exploited for anomaly detection, data imputation and faulty records. Finally, an interesting future direction is related to the choice of the appropriate optimisation function to define the boundaries of subspaces automatically.

References

1. Han, J., J. Pei, and M. Kamber, *Data mining: concepts and techniques*. 2011: Elsevier.
2. Sarle, W.S., *Finding Groups in Data: An Introduction to Cluster Analysis*. 1991, JSTOR.
3. Myatt, G.J. and W.P. Johnson, *Making sense of data II: A practical guide to data visualization, advanced data mining methods, and applications*. 2009: John Wiley & Sons.
4. Deza, M.M. and E. Deza, *Distances and similarities in data analysis*, in *Encyclopedia of distances*. 2013, Springer. p. 291-305.
5. Gower, J.C., *A general coefficient of similarity and some of its properties.* Biometrics, 1971: p. 857-871.
6. Ottaway, B., *Mixed data classification in archaeology.* Revue d'Archéométrie, 1981. **5**(1): p. 139-144.
7. Quinlan, J.R., *Induction of decision trees.* Machine learning, 1986. **1**(1): p. 81-106.
8. Stone, C.J., *Classification and regression trees.* Wadsworth International Group, 1984. **8**: p. 452-456.
9. Salzberg, S.L., *C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993.* Machine Learning, 1994. **16**(3): p. 235-240.
10. Cover, T. and P. Hart, *Nearest neighbor pattern classification.* IEEE transactions on information theory, 1967. **13**(1): p. 21-27.
11. Hopfield, J.J., *Neural networks and physical systems with emergent collective computational abilities.* Proceedings of the national academy of sciences, 1982. **79**(8): p. 2554-2558.
12. Vapnik, V., *The nature of statistical learning theory*. 2013: Springer science & business media.
13. John, G.H. and P. Langley. *Estimating continuous distributions in Bayesian classifiers*. in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. 1995. Morgan Kaufmann Publishers Inc.
14. Hu, L.-Y., et al., *The distance function effect on k-nearest neighbor classification for medical datasets.* SpringerPlus, 2016. **5**(1): p. 1304.
15. Chandrasekar, P., et al. *Improving the prediction accuracy of decision tree mining with data preprocessing*. in *Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual*. 2017. IEEE.
16. Pereira, C.L., G.D. Cavalcanti, and T.I. Ren. *A New Heterogeneous Dissimilarity Measure for Data Classification*. in *Tools with Artificial*

Intelligence (ICTAI), 2010 22nd IEEE International Conference on. 2010. IEEE.

17. Jin, R. and H. Liu. *A Novel Approach to Model Generation for Heterogeneous Data Classification*.
18. Hsu, C.-C., Y.-P. Huang, and K.-W. Chang, *Extended Naive Bayes classifier for mixed data.* Expert Systems with Applications, 2008. **35**(3): p. 1080-1083.
19. Li, X. and N. Ye, *A supervised clustering and classification algorithm for mining data with mixed variables.* IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2006. **36**(2): p. 396-406.
20. Sun, Y., F. Karray, and S. Al-Sharhan. *Hybrid soft computing techniques for heterogeneous data classification*. in *Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on*. 2002. IEEE.
21. Frank, A. and A. Asuncion, *UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California.* School of information and computer science, 2010. **213**.
22. *R data sets*: https://vincentarelbundock.github.io/Rdatasets/datasets.html.