International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 2 (2018) pp. 1201-1208

WDset: A Semantic Distance Measure for Imbalanced Datasets

Hemlata Pant[1] and Dr. Reena Srivastava[2]

[1]Research Scholar, School of Engineering [2]Dean, School of Computer Applications

[1,2] Babu Banarasi Das University, 111, Faizabad Road, Atif Vihar, Lucknow, Uttar Pradesh, 226028, India.

ORCIDs:[1] 0000-0001-6025-9303, [2]0000-0002-0641-6520

Abstract

The concept of similarity/dissimilarity can be used to solve classification problems in imbalanced relational datasets where each record is a set of values corresponding to different attributes. Here, each attribute in the dataset has semantics, and therefore, it has a different effect on determination of category of the records. The distance measures that are commonly used to calculate similarity of the records compromise the semantics of the attributes rather than preserving it.

In this paper, we propose a semantic distance measure named WDset which is a set of weighted distance components from two records. Here, individual distance components are calculated and are assigned a weight based on their relevance. Individual distance components preserve the semantics and weight handles the imbalanced nature of the dataset. Experimental tryouts show that the proposed measure is able to identify similar records more appropriately as compared to the existing measures.

Keywords: Distance Measure, Imbalanced Dataset, Semantics, Data Mining, Classification.

INTRODUCTION

Relational classification is one of the important functionalities of data mining that is used by decision support systems for analysis and strategic decision making. Prime examples include granting loan, issuing credit cards, fraud detection, target marketing, customer relationship management, scientific applications, etc. Classification focuses on categorizing record to a category from a predetermined set of categories. For this a classifier need to be built that can further classify the records. One of the major challenges with classification is the condition where the records belonging to various categories are imbalanced i.e. not equally proportional. This is called class imbalance problem. Class imbalance problem is one of the greatest challenges in machine learning and data mining research. It has acquired significant research interest from academics, industries and research teams in recent years [1, 2, 3, 4, 5, 6, 7, 8, 9]. In this paper, we are considering relational datasets with numerical attributes. All other datasets with non-numerical attributes cab be easily converted to numeric form.

The concept of similarity/dissimilarity can be used to solve classification problems in imbalanced relational datasets. According to this concept similar records belong to the same category and dissimilar records belong to different categories. In a relational dataset, each record is a set of values corresponding to different attributes. Each attribute has semantics, and therefore, it has a different effect on determination of category of the records. Therefore, while calculating the similarity the semantics of the attributes must be preserved for accurate classification. Various measures are available in literature [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26], to compute semantic similarity, but most of them are not suitable for relational imbalanced datasets. The measures that can be used for relational imbalanced datasets are Manhattan [13, 14, 15, 16], Euclidean [12, 13], Weighted Euclidean [16, 17], Minkowski [16, 17] and Dset [19]. However, except Dset all the measures try to combine (in some way) the distance components, and therefore, are not able to preserve the semantics of individual attributes and in turn are not able to determine semantic similarity between the records. Dset preserves semantics of individual attributes but it does not focus on class imbalance problem. Hence, these measures fail to efficiently capture the similarity of records when dataset is imbalanced and a class or classes have significantly more records than that of other classes of the dataset.

In this paper, we propose a distance measure named **WDset.** The proposed distance measure is a semantic distance measure which preserves the semantics of the attributes while determining the distance between the records and it performs well with imbalanced datasets.

Next section discusses the related work done in this area.

RELATED WORK

The distance measures proposed in the literature that can be used for identifying similarity of the records are summarized as below. Firstly, the measure from geometry are described that can be used to determine distance between two records if the records with 'd' attributes are mapped to a d-dimensional space.

The Manhattan distance measure proposed by Minkowski was initially originated as taxicab distance measure with taxicab geometry [14]. This measure was initially defined for R^2 space. For R^2 space the Manhattan distance between two points 'x' and 'y' is calculated as

$$Dist(x, y) = | x_1 - y_1 | + | x_2 - y_2 |$$

where, $x(x_1, x_2)$ and $y(y_1, y_2)$ are two points in the R^2 space.

Later on, the Manhattan distance measure was generalized for R^d space [13,14,15,16], where the Manhattan distance between two points 'x' and 'y' in R^d space is calculated as

$$Dist(x, y) = \sum_{i=1 \text{ to } d} | x_i - y_i |$$

where, $x(x_1, x_2, \ldots, x_d)$ and $y(y_1, y_2, \ldots, y_d)$ are two points in the R^d space.

The Euclidean distance measure gives the distance between two points that can be measured with the help of a ruler [12, 13]. This measure was initially defined for R^2 space and the distance between two points 'x' and 'y' is calculated as

$$Dist(x, y) = \sqrt{(| x_1 - y_1 |^2 + | x_2 - y_2 |^2)}$$

where, $x(x_1, x_2)$ and $y(y_1, y_2)$ are two points in the R^2 space. Later on, a generalized form of Euclidean distance measure was proposed for R^d space with a norm analogous to that defined in R^d space [15-16]. For R^d space the Euclidean distance between two points 'x' and 'y' is calculated as

$$Dist(x, y) = \sqrt{(\sum_{i=1 \text{ to } d} | x_i - y_i |^2)}$$

where, $x(x_1, x_2, \ldots, x_d)$ and $y(y_1, y_2, \ldots, y_d)$ are two points in the R^d space.

Weighted Euclidean distance measures were proposed to emphasize the difference in the predictive importance of the attributes [16, 17]. In the weighted distance measures, individual dimensions are assigned a weight which is based on their predictive importance. This measure calculates the distance between two points 'x' and 'y' as:

$$Dist(x, y) = \sqrt{\sum_{i=1 \text{ to } d} w_i * | x_i - y_i |^2}$$

where, $x(x_1, x_2, \ldots, x_d)$ and $y(y_1, y_2, \ldots, y_d)$ are two points in the R^d space and w_i is the weight assigned to the i^{th} dimension.

The Minkowski distance measure is a metric in a normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance [16, 17]. For R^d space the Minkowski distance of order 'p' between two points 'x' and 'y' is calculated as

$$Dist(x, y) = (\sum_{i=1 \text{ to } d} | x_i - y_i |^p)^{1/p}$$

where, $x(x_1, x_2, \ldots, x_d)$ and $y(y_1, y_2, \ldots, y_d)$ are two points in the R^d space.

Wang et al. proposed an adaptive distance measure [18]. The adaptive distance measures have a property that they adapt according to the record from which distance is determined. For each record 'x' in R a radius 'r_x' is calculated where, 'r_x' is the radius of the largest sphere centered at the record 'x' such that all the records coming within the sphere have the class label same as that of the record 'x'. The radius 'r_x' is calculated as

$$r_x = \min_{y:\ell y=\ell x} Dist(x, y) - €$$

where, $€ > 0$ is an arbitrarily small number and Dist(x, y) is either the Euclidean distance or the Manhattan distance. 'ℓx' and 'ℓy' are class labels associated with the records 'x' and 'y' respectively. Now, the distance between any record 'z' and the record 'x' is calculated as

$$Dist_{new}(z, x) = Dist(z, x) / r_x$$

As calculation of 'r_x' for every record 'x' is a compute intensive process, this method is not efficient for large databases.

In 2009, Reena Srivastava et al. proposed distance measure Dset that preserves the semantics of each attribute while calculating the distance between the records [19]. The distance Dset is an ordered set of distance components one from each attribute. The distance Dset between two records $x(x_1, x_2, \ldots, x_d)$ and $y(y_1, y_2, \ldots, y_d)$ is calculated as

$$Dset = \{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_d - y_d|\}$$

In 2011, Dou Hao et. al. proposed an approach for calculating semantic similarity between words using Word Net [20]. They presented a new edge-counting based method using Word Net to compute the similarity. This method achieves a similarity that perfectly fits with human rating and effectively simulates the human thought process that people prefer to consider more differences when the semantic distance between two words is closer, and vice versa.

In 2014, Zhang Zuo et. al. proposed similarity search of human behavior processes using Extended Linked Data Semantic Distance [21]. They proposed a novel similarity search method that, given a particular behavior, searches for a similar behavior. This method extends the semantic distance calculation method Linked Data Semantic Distance (LDSD) and applies it to human behavior distance calculations.

In 2015 Juhua Hu et.al. proposed pairwised specific distance learning from Physical Linkages [22]. They develop a pairwised specific distance (PSD) approach that exploits the structures of physical linkages and in particular captures the key observations that nonmetric and clique linkages imply the appearance of different or unique semantics, respectively.

In 2016, Tom De Nies et. al. proposed a distance based approach for semantic dissimilarity in knowledge graphs [23]. They introduced a distance-based approach Normalized Semantic Web Distance (NSWD) for measuring the semantic dissimilarity between two concepts in knowledge graph which extends the idea of the Normalized Web Distance, which is utilized to determine the dissimilarity between two textural terms, and utilizes additional semantic properties of nodes in a knowledge graph.

In 2016, Liviu Sebastian Matei et. al. proposed a document semantic distance based on the time series model [24]. They proposed an approach, in which both the documents and the search query are represented as a time series of words. As a consequence, the document is no longer seen as a `bag of words' but as a sequence of words in which the order is essential. They used a well known algorithm in time series, dynamic time warping (DTW), in order to compute the distance between a search query and a document or between two documents. They enhanced the algorithm to

take into consideration the semantic context by using WordNet when computing the distance between words.

In 2016, Badr Hssina et .al. proposed evaluation of semantic similarity using vector space model based on textual corpus [25]. They created a semantic similarity calculation system between text documents to contribute to their semantic clustering.

In 2017, Sultan Alfarhood et .al. propose an approach, called Propagated Linked Data Semantic Distance (PLDSD) [26]. They employed an all-pair shortest path algorithm, the well-known Floyd-Warshall algorithm, to expand semantic distance calculations beyond resources that are just one or two links away.

As mentioned above various measures are available in literature to compute semantic similarity, but most of them are not suitable for relational imbalanced datasets. Dset preserves semantics of individual attributes but it does not focus on class imbalance problem. Hence, the existing measures fail to efficiently capture the similarity of records when dataset is imbalanced and a class or classes have significantly more records than that of other classes of the dataset. Therefore, semantic distance measure for imbalanced dataset is still an issue that needs to be solved.

In the next section, we discuss the proposed distance measure WDset.

WDset

As discussed earlier semantics is associated with each attribute of a relation and is worth preserving while determining the distance between records. Dset, [19], preserves semantics by keeping track of individual distance components; however, for class imbalance problem, importance must be given to the attributes that are more relevant for the identification of imbalanced categories. This can be done by assigning weights to the individual distance components calculated from various attributes according to their relevance.

This paper proposes a distance measure WDset that is a set of weighted distance components from individual attributes. WDset extends Dset [19], by assigning weight to each distance component. The distance WDset is calculated as

$$WDset = \{w_1 * |x_1 - y_1|, w_2 * |x_2 - y_2|, \ldots, w_d * |x_d - y_d|\}$$

Where, w_1 ,$w_2 \ldots\ldots w_d$ are the weights assigned to each attribute as per its relevance. For calculating weights, we have used the normalized value of the feature selection measure Mindex_IB [27]. Here the weights w_1, $w_2 \ldots\ldots w_d$ are calculated as $w_i = Mindex\_IB_i / W_f$, where, $Mindex\_IB_i$ is the normalized value of relevance generated by the feature selection measure Mindex_IB for i^{th} attribute and W_f is the weight factor. W_f may vary from dataset to dataset to deal with the different level of imbalance and can be adjusted to find semantically similar records in a better way.

Arithmetic and Relational Operations on the Proposed Distance Measure

This subsection defines the arithmetic and relational operations on the proposed distance measure.

Let $WDset_1 = \{w_1ds_{11}, w_2ds_{12}, \ldots, w_dds_{1d}\}$ and $WDset_2 = \{w_1ds_{21}, w_2ds_{22}, \ldots, w_dds_{2d}\}$ be the two WDset distances. Then

1. $WDset_1 + WDset_2 = \{| w_1ds_{11} + w_1ds_{21} |, | w_2ds_{12} + w_2ds_{22} |, \ldots, | w_dds_{1d} + w_dds_{2d} |\}$

2. $WDset_1 - WDset_2 = \{| w_1ds_{11} - w_1ds_{21} |, | w_2ds_{12} - w_2ds_{22} |, \ldots, | w_dds_{1d} - w_dds_{2d} |\}$

3. $WDset_1 = WDset_2$, if for i=1 to d $w_ids_{1i} = w_ids_{2i}$

4. $WDset_1 \leq WDset_2$, if for i=1 to d $w_ids_{1i} \leq w_ids_{2i}$

5. $WDset_1 < WDset_2$, if $WDset_1 \leq WDset_2$ and $\exists\, i \in \{1, \ldots, d\}$ such that $w_ids_{1i} < w_ids_{2i}$

6. $WDset_1 \geq WDset_2$, if for i=1 to d $w_ids_{1i} \geq w_ids_{2i}$

7. $WDset_1 > WDset_2$, if $WDset_1 \geq WDset_2$ and $\exists\, i \in \{1, \ldots, d\}$ such that $w_ids_{1i} > w_ids_{2i}$

Distance Axioms

Any distance measure should satisfy the distance axioms of equal self similarity, minimality, symmetry, and triangle inequality. WDset fulfill these axioms as discussed here. Let there be three records $A(a_1, a_2, \ldots, a_d)$, $B(b_1, b_2, \ldots, b_d)$ and $C(c_1, c_2, \ldots, c_d)$ from a relation with d attributes.

1. Equal Self Similarity: dist(A,A) = dist(B,B).

$$WDset(A,A) = \{| a_1 - a_1 | * w_1, | a_2 - a_2 | * w_2, \ldots, | a_d - a_d | * w_d\} = \{0, 0, \ldots, 0\}.$$

In the same way,

$$WDset(B,B) = \{| b_1 - b_1 | * w_1, | b_2 - b_2 | * w_2, \ldots, | b_d - b_d | * w_d\} = \{0, 0, \ldots, 0\}.$$

Thus, WDset(A,A) = WDset(B,B).

2. Minimality: dist(A,B) > dist(A,A).

WDset(A,A) = {0, 0, . . . , 0} and WDset(A,B) = {| $a_1 - b_1$ |*w_1, | $a_2 - b_2$ |*w_2, . . . , |$a_d - b_d$|*w_d}.

In the distance Dset(A,B) when most of the distance components are zero, at least one of the components will be greater than zero. This is because A and B are two distinct records. So,

WDset(A,B) > WDset(A,A).

3. Symmetry: dist(A,B) = dist(B,A).

WDset(A,B) = {| $a_1 - b_1$ |*w_1, | $a_2 - b_2$ |*w_2, . . . , | $a_d - b_d$ |*w_d}

WDset(B,A) = {| $b_1 - a_1$ |*w_1, | $b_2 - a_2$ |*w_2, . . . , | $b_d - a_d$ |*w_d}

= {| $a_1 - b_1$ |*w_1, | $a_2 - b_2$ |*w_2, . . . , | $a_d - b_d$ |*w_d}

= WDset(A,B).

Thus, Dset(A,B) = Dset(B,A).

4. Triangle Inequality: dist(A,B) + dist(B,C) ≥ dist(A,C).

WDset(A,C) = {| $a_1 - c_1$ |*w_1, | $a_2 - c_2$ |*w_2, . . . , | $a_d - c_d$ |*w_d}

WDset(A,B) +W Dset(B,C) = {| $a_1 - b_1$ | *w_1+ | $b_1 - c_1$ |*w_1, | $a_2 - b_2$ |*w_2 + | $b_2 - c_2$ |*w_2, . . . ,| $a_d - b_d$ | *w_d+ | $b_d - c_d$ |*w_d}

To prove the triangle inequality we have to prove that

| $a_i - b_i$ | *w_i+ | $b_i - c_i$ |*w_i≥| $a_i - c_i$ |*w_i, for i=1 to d.

To prove this let us map the three records A, B and C to three points in the d-dimensional Euclidean space. Here, three possibilities exist. First, a_i lies between b_i and c_i. Second, bi lies between a_i and c_i and in the third case, c_i lies between a_i and b_i. Let us take the first case when a_i lies between b_i and c_i. Here,

| $a_i - b_i$ |*w_i + | $b_i - c_i$ |*w_i

= | $a_i - b_i$ |*w_i + | $b_i - a_i$ | *w_i+ | $a_i - c_i$ |*w_i

= 2* | $a_i - b_i$ |*w_i + | $a_i - c_i$ |*w_i

≥ | $a_i - c_i$ |*w_i

In the second case when bi lies between a_i and c_i. Here

| $a_i - b_i$ | *w_i+ | $b_i - c_i$ |*w_i

= | $a_i - c_i$ |*w_i

Finally, in the third case when c_i lies between a_i and b_i.

| $a_i - b_i$ | *w_i+ | $b_i - c_i$ |*w_i

= | $a_i - c_i$ |*w_i + | $c_i - b_i$ |*w_i + | $b_i - c_i$ |*w_i

= | $a_i - c_i$ |*w_i + | $b_i - c_i$ | *w_i+ | $b_i - c_i$ |*w_i

= | $a_i - c_i$ |*w_i +2* | $b_i - c_i$ |*w_i

≥ | $a_i - c_i$ |*w_i

Hence, the distance measure Dset satisfies the triangle inequality axiom as well as all the distance axioms.

Table 1: Dataset Used

Name	Total Records	Features	#Class	#Instances Per Class
Ecoli	336	7	8	143, 77, 52, 35, 20, 5, 2, 2
Ionosphere	351	34	2	126, 225
Pima Indians Diabetes	768	8	2	268, 500
Glass Identification	214	10	6	70, 76, 17, 13, 9, 29
Fertility	100	9	2	88, 12

EXPERIMENTAL BACKGROUNDS

Experiments are based on the concept of Micro-clusters [28]. A micro-cluster is a small cluster (enclosing some records) in a d-dimensional Euclidean space where, d is the number of the attributes used to identify the micro-clusters. Records from a relation are mapped to a d-dimensional space and considered as points. Micro-clustering is based on the concept that, points close to each other in the space are similar to each other, and thus, belong to the same category. To check similarity of records in a micro-cluster set, entropy is used. Entropy is a measure of similarity of the records in a micro-cluster within a micro-cluster set. While measuring entropy, probability of records belonging to different categories is considered. Since category is involved, entropy measures the semantic similarity of the records. If the entropy measure indicates that the records in the micro-clusters within a micro-cluster set are semantically similar, then this means that the distance measure used to determine the micro-clusters is able to capture semantic similarity among records. The smaller the entropy the more appropriate the distance measure is. However, if the entropy value is very small or zero and the number of micro-clusters approach the number of records, Then the small value of entropy will not be able to justify the purpose as each micro-cluster contains only one or two records. Therefore, a tradeoff between the number of micro-clusters and the entropy value should be analyzed.

To compare the proposed distance measure WDset with the existing distance measures, micro-clusters are identified from the dataset using these distance measures. The micro-clusters are identified for the value of K=128 (as this is the value considered by MINDEX_IB [27]. After determining the micro-clusters entropy of the micro-cluster set is calculated and compared.

EXPERIMENTAL RESULTS AND DISCUSSION

In order to evaluate the performance of the proposed distance measure WDset, the datasets from the UCI repository,

namely, Ecoli, Ionosphere, Pima Indians Diabetes, Glass Identification and Fertility [29], with the details as shown in Table I are used.

Figure 1 shows the results for Ecoli dataset. Here, for Dset, Euclidean and Manhattan distance measures the number of micro-clusters for all the values of weight factor (W_f) is equal to the number of records which means that they are not able to make proper micro-clusters. Moreover, for weighted Euclidean measure also the number of micro-clusters is very high. Therefore, the corresponding low value of entropy for these measures is not significant. With WDset as the value of W_f is increased proper micro-clusters are formed. For Ecoli dataset, the value of W_f =32 where 71 micro-clusters are formed with entropy as 0.56, can be used to determine similarity of the records.

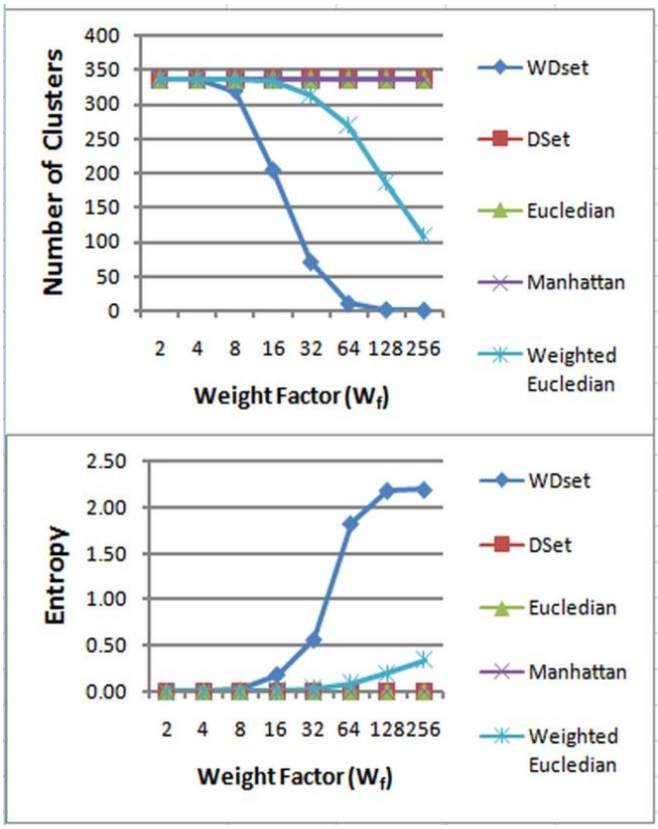

Figure 1: Number of Clusters and Entropy w.r.t. Weight Factor for Ecoli Dataset

Figure 2 shows the results for Ionosphere dataset. Here, for Dset, Euclidean and weighted Euclidean distance measures the number of micro-clusters for all the values of weight factor (W_f) is equal to the number of records. Moreover, with Manhattan distance measure only 1 micro-cluster is formed. Therefore, these measures are not able to make proper micro-clusters. However, with WDset as we increase the value of W_f we are able to get proper micro-clusters. For Inosphere dataset, the value of W_f =64 where 96 micro-clusters are formed with entropy as 0.26, can be used to determine similarity of the records.

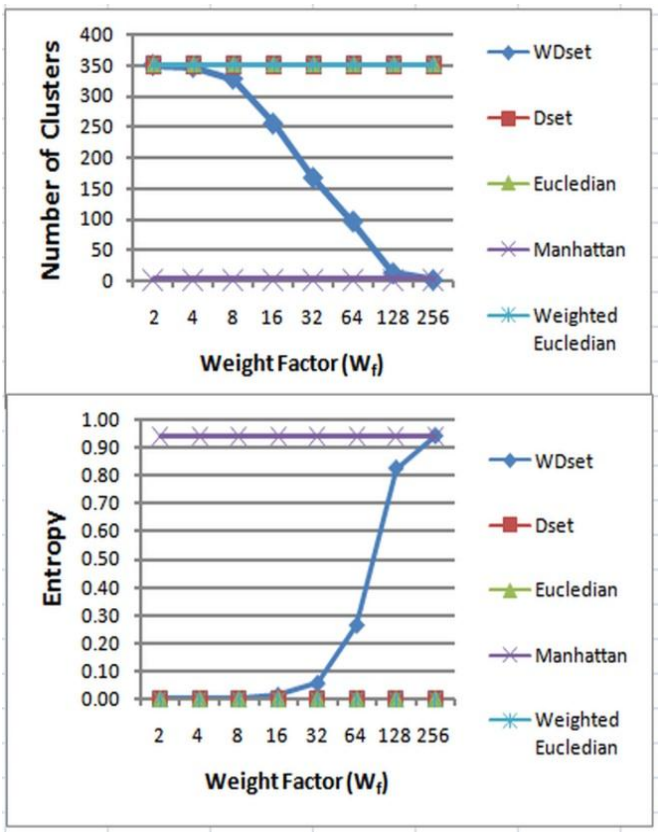

Figure 2: Number of Clusters and Entropy w.r.t. Weight Factor for Ionosphere Dataset

Figure 3 shows the results for Pima Indians Diabetes dataset. Here, for Dset, Euclidean and Manhattan distance measures the number of micro-clusters for all the values of weight factor (W_f) is equal to the number of records which means that they are not able to make proper micro-clusters. For weighted Euclidean measure also the number of micro-clusters is very high. Therefore, the corresponding low value of entropy for these measures is not significant. However, with WDset we are able to get proper micro-clusters. For Pima Indians Diabetes dataset, the value of W_f =32 where 34 micro-clusters are formed with entropy as 0.77, can be used to determine similarity of the records.

Figure 4 shows the results for Glass Identification dataset. Here, for Dset and Euclidean distance measures the number of micro-clusters for all the values of weight factor (W_f) is equal to the number of records and for Manhattan distance measure only 1 micro-cluster is formed. Therefore, these measures are not able to make proper micro-clusters. For weighted Euclidean distance measure the number of micro-clusters is low but the entropy is very high. However, with WDset we are able to get proper micro-clusters. For Glass Identification dataset, the value of W_f =16 where 56 micro-clusters are formed with entropy as 0.82, can be used to determine similarity of the records.

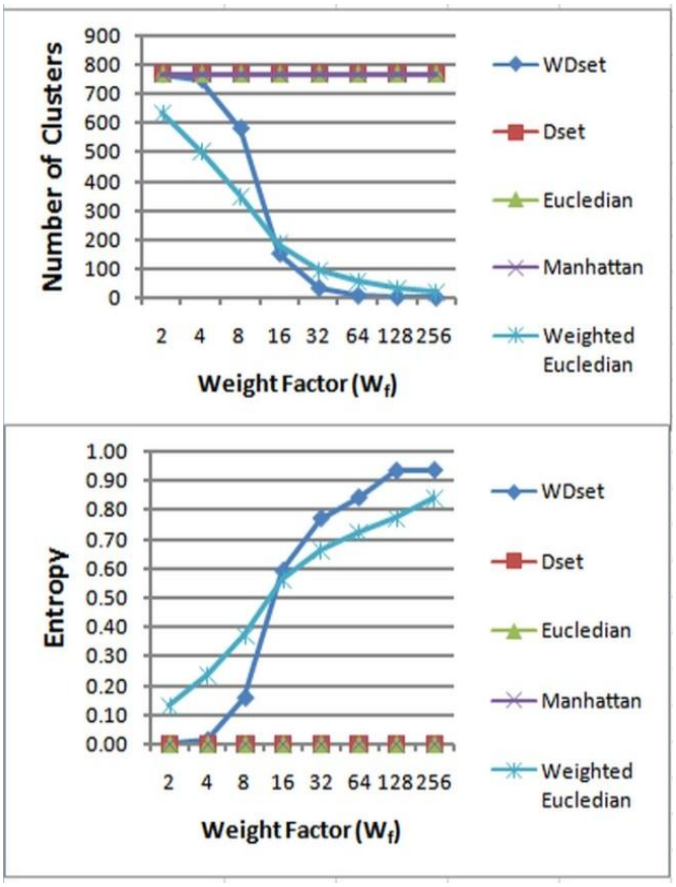

Figure 3: Number of Clusters and Entropy w.r.t. Weight Factor for Pima Indians Diabetes Dataset

Figure 5 shows the results for Fertility dataset. Here, for Dset, Euclidean, Manhattan and Weighted Euclidean distance measures the number of micro-clusters for all the values of weight factor (W_f) is equal to the number of records indicating the fact that they are not able to make proper micro-clusters. However, with WDset we are able to get proper micro-clusters. For Fertility dataset, the value of W_f =256 where 38 micro-clusters are formed with entropy as 0.35, can be used to determine similarity of the records.

The experimental results for the above mentioned five datasets validates that the proposed distance measure WDset determines semantic similarity for imbalanced datasets in a much better way as compared with others.

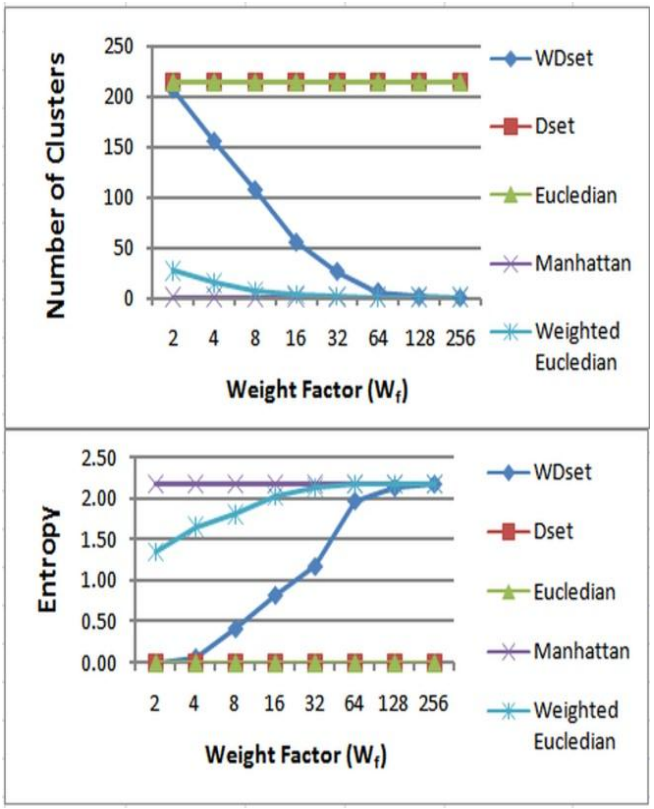

Figure 4: Number of Clusters and Entropy w.r.t. Weight Factor for Glass Identification Dataset

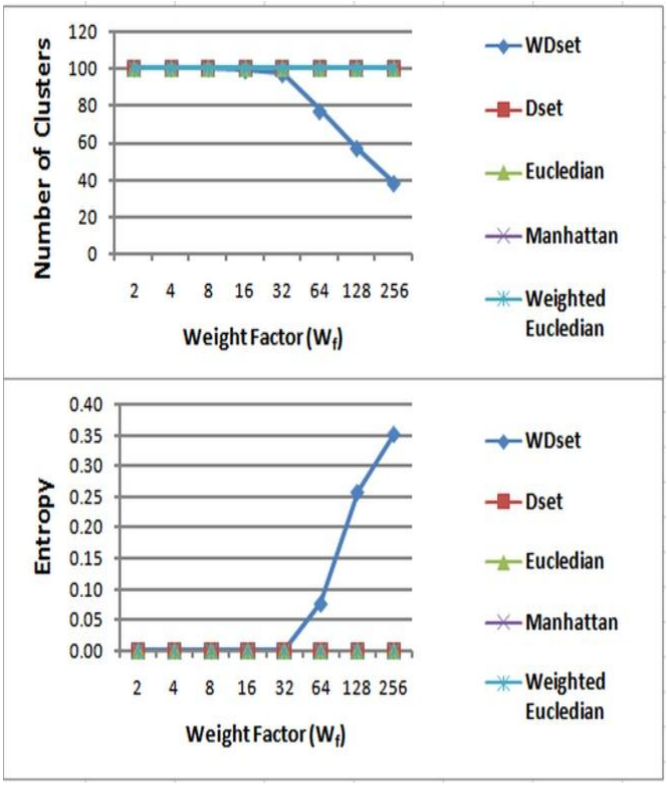

Figure 5: Number of Clusters and Entropy w.r.t. Weight Factor for Fertility Dataset

CONCLUSION

This paper proposes a distance measure WDSet for imbalanced dataset. WDset is a semantic distance measure which preserves the semantics associated with each attribute while measuring the distance between two records in the dataset and also deals with the imbalanced nature of the dataset with the help of weights. Calculation of weights includes a factor W_f that may vary from dataset to dataset to deal with the different level of imbalance and can be adjusted to find semantically similar records in a better way.

Experimental results for the proposed distance measure WDset shows that by adjusting the value of Weight Factor (W_f) we can easily arrive at a value of W_f where the number of micro-clusters is less and the entropy of the micro-cluster set is also low which in turn indicates that semantically similar records were identified.

REFERENCES

[1] M. Kubat, S. Matwin. "Addressing the curse of imbalanced training sets: one sided selection". Proc.14th International Conference on Machine Learning, pp. 179–186, 1997.

[2] Alexander Liu, J. Ghosh, C. Martin. "Generative oversampling for mining imbalanced datasets". Proc. International Conference on Data Mining, Las vegas, Nevada, USA CSREA Press. pp. 66-72, 2007.

[3] Chien-I Lee, Cheng-Jung Tsai, Tong-Qin Wu et al. "An approach to mining the multi-relational imbalanced database". Expert Systems with Applications. Vol 34:3021-3032, 2008.

[4] H. Guo, H. L. Viktor. "Mining imbalanced classes in multirelational classification". Proc. 6th Multi-Relational Data Mining Workshop (PKDD/MRDM'07), in conjunction with 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'07), Warsaw, Poland, pp. 46-57, 2007.

[5] Y. Murphey, H. Wang, G. Ou et al. "OAHO: an effective algorithm for multi-class learning from imbalanced data". Proc. International Joint Conference on Neural Networks (IJCNN), pp. 406–411, 2007.

[6] A.S. Ghanem, S. Venkatesh, G. West. "Learning in Imbalanced Relational Data". Proc. 19th International Conference on Pattern Recognition, Tampa, Florida, USA. IEEE Computer Society, pp. 1-4. 2008.

[7] S.J. Yen and Y.S. Lee. "Cluster-based under-sampling approaches for imbalanced data distributions". Expert Systems with Applications.2009; Vol. 36, pp. 5718–5727, 2009.

[8] A.S. Ghanem, S. Venkatesh, G. West. "Multiclass pattern classification in imbalanced data". Proc. International Conference on Pattern Recognition, Istanbul, Turkey, pp. 2881-2884, 2010.

[9] P. Jeatrakul, K.W. Wong. "Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm". Proc. Annual International Joint Conference on Neural Networks, IJCNN, Brisbane, Australia, pp. 1-8,2012

[10] C. Domeniconi, J. Peng, D. Gunopulos. "Locally adaptive metric nearest-neighbor classification". IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 9, 2, pp. 1281–1285, 2002.

[11] J.H. Friedman. "Flexible metric nearest neighbor classification". Technical Report No. 133, Stanford University, 1994.

[12] E.F. Krause. "Taxicab Geometry: An Adventure in Non-Euclidean Geometry". Dover, Mineola, NY, 1986.

[13] E. Kreyszig. "Introductory functional analysis with applications," Wiley, India, 2007.

[14] C. Reinhardt. "Taxi cab geometry: History and applications". The Montana Mathematics Enthusiast 12, 1 (2005), pp.38–64, 2005.

[15] R.N. Shepard. "Toward a universal law of generalization for psychological science". Science 237, 4820 (1987), pp.1317–1323, 1987.

[16] M. Steyvers. "Multidimensional scaling". In Encyclopedia of Cognitive Science (2002), Nature, 2002.

[17] J. Han, M. Kambar. "Data Mining Concepts and Techniques". First Indian Reprint ed. Morgan Kaufmann, 2001.

[18] J. Wang, P. Neskovic and L.N. Cooper. "Improving nearest neighbour rule with a simple adaptive distance measure". Pattern Recognition Letters 28, pp. 207–213, 2007.

[19] R. Srivastava, M.M. Gore. "Dset: Semantic Distance Measure" In Proceedings of 3rd International Conference on Information Processing (ICIP), Bangalore, India, pp. 419-428, 2009.

[20] D. Hao, W. Zuo, T. Peng et al. "An approach for calculating semantic similarity between words using WordNet". Proc. Second International Conference on Digital Manufacturing & Automation, Zhangjiajie, Hunan, China, pp. 177-180, 2011.

[21] Z. Zuo, H. Huang, K. Kawagoe. "Similarity search of human behavior processes using extended linked data semantic distance". Proc. 25th International Workshop on Database and Expert Systems Applications, Munich, Germany, pp. 178-182, 2014.

[22] H.U. Juhua, Z. De-chuan, W.U. Xintao et al. **"**Pairwised specific distance learning from physical linkages". ACM Transactions on Knowledge Discovery from Data, Vol. 9, No. 3, Article 20, 2015.

[23] T. De Nies, C. Beecks, F. Godin et al. "A Distance-based approach for semantic dissimilarity in knowledge graphs". Proc. IEEE Tenth International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, pp. 254-257, 2016.

[24] L.S. Matei, S.T. Matu. "Document semantic distance based on the time series model". Proc. 15th RoEduNet Conference: Networking in Education and Research, Bucharest, Romania, 2016.

[25] B. Hssina, B. Bouikhalene, A. Merbouha. "Evaluation of semantic similarity using vector space model based on yextual corpus". Proc. 13th International Conference on Computer Graphics,Imaging and Visualization (CGiV), Beni Mellal, Morocco, 2016.

[26] S. Alfarhood, K. Labille, S. Gauch. "PLDSD: Propagated linked data semantic distance". Proc. 26th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Poznan, Poland, pp.278-283, 2017.

[27] H. Pant, R. Srivastava. "MINDEX_IB: A Feature Selection method For Imbalanced Dataset". International Journal of Advanced Research in Computer and Communication Engineering, Volume V, Issue IV, ISSN 2278-1021, pp. 766-769, 2016.

[28] C.C. Aggarwal, J. Han, J. Wang et al. "P. S. A framework for clustering evolving data streams". Proc. 29th International Conference on Very Large Data Bases VLDB, Morgan Kaufmann, pp. 81–92,2003

[29] http://archive.ics.uci.edu