Correlation and Probability Based Similarity Measure for Detecting Outliers in Categorical Data

Roy Thomas, J.E.Judith

Abstract: Determining the similarity or distance among data objects is an important part in many research fields such as statistics, data mining, machine learning etc. There are many measures available in the literature to define the distance between two numerical data objects. It is difficult to define such a metric to measure the similarity between two categorical data objects since categorical data objects are not ordered. Only a few distance measures are available in the literature to find the similarities among categorical data objects. This paper presents a comparative evaluation of various similarity measures for categorical data and also introduces a novel similarity measure for categorical data based on occurrence frequency and correlation. We evaluated the performance of these similarity measures in the context of outlier detection task in data mining using real world data sets. Experimental results show that the proposed similarity measure outperform the existing similarity measures to detect outliers in categorical datasets. The performances are evaluated in the context of outlier detection task in data mining.

Keywords: Categorical, Correlation, Outlier, Similarity.

I. INTRODUCTION

Measuring the distance between two data objects is one of the key requirements for many data mining tasks such as clustering, classification, neighborhood based outlier detection. The success of the algorithms used for these data mining tasks heavily depends on the choice of the distance measure. The dissimilarity or distance between two objects is a numerical measure of the degree to which the two data objects are different and the similarity between two objects is a numerical measure of the degree to which the two objects are alike. The absolute difference between the numerical values representing the quantitative property may be considered as a similarity measure for continuous or numerical data objects. Minkowski distance of order two, which is called the Euclidean distance, is one of the most widely used distance measures for numerical data. Measuring similarity between categorical data objects is not so easy as in the case of continuous valued data objects since a numerical value representing a qualitative trait is not available for comparing the categorical data objects. Categorical data objects are characterized by some qualitative traits. Categorical variables can be classified into ordinal and nominal, based on the existence of inherent ordering among the values taken by the variable. A nominal variable is one that has two or more values, but there is no intrinsic ordering among the values. For example, color is a nominal variable having different values like green, red, white etc. and there is no intrinsic ordering among these values. An ordinal variable has values with clear ordering among them. The grades obtained by the students in an examination having the values average, good, very good, outstanding is an example of ordinal variable. Here the similarity between the values ‘very good’ and ‘outstanding’ is higher than the similarity between ‘average’ and ‘outstanding’. In this paper, by categorical data we mainly refer to the nominal data in which there is no inherent ordering among the values. Traditionally the similarity between two nominal data objects are calculated based on the number of matches or mismatches between the data objects. Data-driven similarity measures for categorical data make use of the frequency distribution and co-occurrence statistics to define the degree of similarity between the data objects.

II. RELATED WORK

Distance measure is an important metric in data mining and has a wide range of applications in many fields. It is often needed in data mining to identify how far an object is from other objects in the data set or how much similarities exist among the data objects. A number of clustering and outlier detection algorithms in data mining are based on nearest neighborhood. These algorithms require a distance measure between one data object and another data object in the dataset. A large number of distance measures are available in the literature for computing the distance in numerical data objects such as Minkowski distance. It is also important for determining the distance between categorical data objects as in numerical or continuous data objects. But there is no such generalized measure exists to determine the distance or dissimilarity among categorical data objects due to the dependency of domain knowledge in determining the similarities in categorical data. Han et al[1] detailed the various similarity measures for numerical as well as nominal data including binary attributes and ordinal attributes. They presented some of the similarity and dissimilarity measures in the literature such as Euclidean distances, Manhattan distances, Minkowski distances, Jaccard coefficient and cosine similarity. Aggarval [2] categorized the similarity measures into two primary classes – methods based on the aggregate statistical properties of the data and methods based on the statistical neighborhoods of the data points. Boriah et al [3] accumulated many similarity measures in the literature, proposed some new measures derived from the existing measures and evaluated the performance of these measures on different data sets.

III. SIMILARITY MEASURES

Data objects in the real world are denoted by a set of attributes. Some of these attributes may represent the
Correlation and Probability Based Similarity Measure for Detecting Outliers in Categorical Data

A. Overlap

Overlap similarity measure is one of the most common similarity measures used in the literature due to its simplicity. In overlap similarity measure, a similarity value of 1 is assigned for matches and 0 is assigned for mismatches. The overlap similarity between two values x_k and y_k of attribute A_k is defined as

\[ S_k(x_k, y_k) = 1 \quad \text{if} \quad x_k = y_k \]
\[ S_k(x_k, y_k) = 0 \quad \text{if} \quad x_k \neq y_k \]

B. Eskin

Eskin is another similarity measure in which the maximum value for similarity is assigned to all matches and varying weights are assigned to mismatches. The weight assigned by this measure is based on the number of values nk taken by an attribute. This measure assigns less weight to mismatches that occur on attributes that contain fewer values and assigns the maximum value when the values in the attribute are all unique. The Eskin similarity between two values is computed as

\[ S_k(x_k, y_k) = 1 \quad \text{if} \quad x_k = y_k \]
\[ S_k(x_k, y_k) = \frac{n^2}{n^2 + 2} \quad \text{if} \quad x_k \neq y_k \]

C. IOF

The inverse occurrence frequency(IOF) measure also assigns the maximum value 1 to all matches. Similarity on mismatches is assigned based on the frequency of values f(x_k) and f(y_k). This measure assigns lower value to mismatches on more frequent values. The highest value for mismatches is attained when the values occur only once in the data set and the minimum value is attained when they are the only two values for the attribute and each occur equal times. The mathematical formulae for computing IOF similarity measure is given as,

\[ S_k(x_k, y_k) = 1 \quad \text{if} \quad x_k = y_k \]
\[ S_k(x_k, y_k) = \frac{1}{1 + \log f(x_k) \log f(y_k)} \quad \text{if} \quad x_k \neq y_k \]

D. OF

The occurrence frequency measure assigns the same value for matches as given by inverse occurrence frequency, but assigns reverse weighting of the IOF measure for mismatches. The OF measure assigns lower similarity to less frequent values and higher similarity to more frequent values. The lowest value is assigned when the measuring values occur only once in the data set and the highest value is assigned when these measuring values are the only two values and each having the same frequency. Mathematically, the OF measure can be computed using the formulae given below.

\[ S_k(x_k, y_k) = 1 \quad \text{if} \quad x_k = y_k \]
\[ S_k(x_k, y_k) = \frac{1}{1 + \log \frac{x_k}{y_k} \log \frac{y_k}{x_k}} \quad \text{if} \quad x_k \neq y_k \]

E. Lin
The Lin similarity measure gives varying weights to both matches and mismatches. The Lin measure assigns higher weight to frequent values, and lower weight to infrequent values. The similarity $S_k(X_k, Y_k)$ is assigned the minimum value for a match when $X_k$ occurs only once in the data set and the maximum value is assigned when $X_k$ has only one value. The minimum value for a mismatch is obtained when each value occurs only once and the maximum value is obtained when $X_k$ and $Y_k$ are the only values for this attribute and each occur equal times. The formulae for this measure are,

$$S_k(X_k, Y_k) = 2\log P_k(X_k) \quad \text{if } X_k = Y_k$$
$$S_k(X_k, Y_k) = 2\log(P_k(X_k) + P_k(Y_k)) \quad \text{if } X_k \neq Y_k$$

F. Goodall4

Goodall4 is one of the variants of the Goodall similarity measure. The Goodall4 measure assigns a value of 0 for all mismatches and assigns weight for matches depending on the frequency of the values. The maximum value for matches in Goodall4 similarity measure $S_k(X_k, Y_k)$ is assigned when $X_k$ is the only value for the $k^{th}$ attribute and minimum value is assigned when $X_k$ occurs only once. The mathematical formulae for this measure are given below.

$$S_k(X_k, Y_k) = P_k^2(X_k) \quad \text{if } X_k = Y_k$$
$$S_k(X_k, Y_k) = 0 \quad \text{if } X_k \neq Y_k$$

G. Correlation Weighted Probability and Level Based Similarity (CPL Similarity)

In this proposed CPL similarity measure correlation between an attribute and the class attribute, $\text{Corr}(A_k, A_{class})$, is used as the weight for the similarity. This measure also considers $n_k$, the number of values taken by each attribute and the probability of occurrence of these values. This method assigns different weights to mismatches based on the number of levels in each attribute and the probability of occurring each values $P_k(X_k)$ and $P_k(Y_k)$. It assigns more weight to mismatches on attributes having many levels. This measure also assigns higher weight for less frequent values.

We used Goodman and Kruskal’s tau (GKtau) measure[4] based correlation between an attribute $A_k$ and the class attribute $A_{class}$ for determining the weight for per attribute similarity of $A_k$. The main reason for selecting this measure as the correlation ratio is that GKtau is independent of the order among the categorical data. Also, GKtau measure is a good measure for measuring the strength of association between categorical variables. The results given by GKtau measure for predicting $x$ from $y$ and for predicting $y$ from $x$ are generally different values[5]. The proposed CPL similarity measure uses the association of each attribute with the class variable only.

$$S_k(X_k, Y_k) = \text{Corr}(A_k, A_{class}) \quad \text{if } X_k = Y_k$$
$$S_k(X_k, Y_k) = \frac{\text{Corr}(A_k, A_{class})}{n_k \star (P_k(X_k) - P_k(Y_k))^2 + 2} \quad \text{if } X_k \neq Y_k$$

IV. OUTLIER DETECTION METHODS

Outlier detection is aimed at finding infrequent data objects that are not conformable with the mainstream characteristics of the majority of data objects in the data set[6]. As shown in Fig. 1, an outlier is a data object that is not clustered with the most of the remaining data objects in the data set. Outlier detection may be defined as the discovery of dissimilar data objects from a data set containing a large collection of objects in which the characteristics of a small number of objects are not in accordance with the characteristics of the majority of data objects in the data set.

Fig. 1. Objects outside the ellipse are Outliers and objects inside the ellipse are normal objects

Outlier detection has become an important activity in data science as it provides vital information in a variety of domains[7]. Outlier detection methods can be applied to detect unauthorized access to computer network and servers, to find spam messages and mails, to detect diseases in medical data analysis, to find forged credit card transactions in banking etc.

Depending on the availability of class labels, there are supervised, semi-supervised, and unsupervised outlier detection methods [8]. Outlier detection methods can also be grouped into different categories[9] like statistical methods, proximity based methods, classification based methods and clustering based methods, depending on the assumptions made on outliers versus normal data.

A. Statistical Methods

Statistical outlier detection methods assume that normal data objects follow a statistical model, and data objects that do not follow this statistical model are considered as outliers. A statistical test is made to divide the data objects into normal data objects and outliers. During the test the normal objects occur in high probability region of the model and outliers occur in low probability regions. Statistical models are mostly suitable for numerical data sets and are not appropriate for categorical data sets without any numerical transformation. It is difficult to obtain the suitable numerical transformation without loss of information and the conversion increases the preprocessing complexity. This limits the applicability of statistical techniques in detecting outliers.

B. Classification Based Methods

Classification based approaches are used to find outliers in labeled data sets.
These methods use learned classifiers from the data set containing labeled data objects to classify the data objects into normal data objects and outliers. These methods use two phases – the training phase and the testing phase. During the training phase the classifier is built using the labeled data objects and in the testing phase the data objects are classified into normal objects and outliers.

C. Clustering Based Methods

Clustering based methods are used when the class labels are not available. These methods divide the data set into different clusters, where each cluster contains objects having similar characteristics. Data objects in large and dense clusters are considered as normal data objects and other data objects are considered as outliers. Hence outliers can be data objects in small or sparse clusters, or can be data objects that do not belong to any of the clusters.

D. Proximity Based Methods

Proximity based methods are developed on the assumption that the distance from an outlier to its nearest neighbors is higher than the distance from a normal data object to its nearest neighbors. Normal data instances are assumed to form high density regions and outliers form low density regions. In proximity based methods, a similarity measure or a distance measure defined between two data objects is needed to compute the outlier score of each data object. The effectiveness of the methods relies on this measure. Proximity based methods can be used only when such a similarity or distance measure is available to compute the distance between data objects and hence the outlier score. These methods are used to find global or point outliers and are not suitable to detect collective outliers. Proximity based outlier detection methods are further divided into different categories like distance based methods, density based methods etc. In this paper, we evaluate the effectiveness of the proposed correlation weighted probability and level based (CPL) similarity measure to detect outliers in categorical data using proximity based outlier detection methods. A comparative evaluation of the performance of this proposed CPL method with other similarity measures is also presented.

V. EXPERIMENTS

In this section, we describe the experimental settings and the datasets used in Section 5.1 followed by experimental results in Section 5.2.

A. Experimental Settings

Publicly available datasets from the UCI machine learning repository are used for our experiments. The datasets used are ‘Car’ dataset, ‘Census’ dataset and ‘Lenses’ dataset. Some of these datasets contain numerical data types as well as categorical attributes. The numerical attributes are removed from the mixed type datasets for our experiments. Only the categorical attributes and random subsets of tuples from each datasets are used to evaluate the various similarity measures. The random subset of data from each datasets used for the experiments contain only two classes of data, in which the major class is identified as normal class and the other small set is treated as outlier class. The description of the datasets is given in Table 1.

| Dataset   | Instances | Attributes | Attribute type | classes | Normal | Outlier |
|-----------|-----------|------------|----------------|---------|--------|---------|
| Car       | 70        | 6          | Categorical    | 2       | 65     | 5       |
| Census    | 121       | 9          | Categorical    | 2       | 114    | 7       |
| Lenses    | 19        | 4          | Categorical    | 2       | 15     | 4       |

Experiments using these publicly available datasets were conducted in an intel core i3 based laptop using R. Separate experiments were conducted for each similarity measures using different datasets. The original datasets contain multiple classes. Only two classes of data objects were used in the experiment for evaluating the similarity measure using outlier detection technique. Samples for experiments were taken from the dataset randomly in which the majority of the data belongs to the normal class and the remaining small set as outliers. The similarity measures used for the evaluation are Overlap similarity, Eskin similarity, IOF similarity, OF similarity, goodall4 similarity and the proposed CPL similarity. These similarity measures are used as a distance measure to detect outliers using proximity based methods. The relative performance of the similarity measures were evaluated based on the performance of these measures in detecting outliers using k-nearest neighbor classification with the input parameters k and the percentage of outliers.

VI. RESULTS

The evaluation measures used for comparing the performance of different algorithms are accuracy, error rate, precision and recall. The values of these measures obtained from these algorithms using census dataset, lenses dataset and car dataset are tabulated in Table 2, Table 3 and Table 4 respectively. Results obtained from the different algorithms were compared with the actual dataset and evaluated the performance of each similarity measures. Analysis of the results obtained from the various algorithms shows that the performance of Overlap, Eskin, IOF, OF, Lin and goodall4 measures depend on the nature of the dataset and the proposed CPL measure outperforms all these measures.

Fig. 2 shows the evaluation metric accuracy obtained from these similarity measures using different datasets. The CPL measure gives the highest value for all the different datasets used for the evaluation. Fig.3 indicates that the proposed CPL measure has the least error rate for all the three datasets. Similarly Fig. 4 and Fig.5 show that the CPL gives higher values for recall and precision. These indicate that the performance of the similarity measure can be improved by assigning the correlation between each attribute and the class attribute as weight in the similarity computation and also considering the probability of occurrence of each attribute in the dataset. The CPL similarity measure can be used as a better distance measure for categorical data to find
the outliers in categorical datasets using neighborhood based outlier detection.

### Table 2. Performance of similarity measures in detecting outliers using Census dataset

| Measure | Overlap | Eskin | IOF | OF | Lin | Goodall4 | CPL |
|---------|---------|-------|-----|----|-----|----------|-----|
| TP      | 3       | 1     | 1   | 3  | 1   | 2        | 5   |
| TN      | 106     | 107   | 107 | 109| 107 | 108      | 111 |
| FP      | 8       | 7     | 7   | 5  | 7   | 6        | 3   |
| FN      | 4       | 6     | 6   | 4  | 6   | 5        | 2   |
| Accuracy| 0.90083 | 0.89256 | 0.89256 | 0.92562 | 0.89256 | 0.90909 | 0.95868 |
| Error rate| 0.09917 | 0.10744 | 0.10744 | 0.07438 | 0.10744 | 0.09091 | 0.04132 |
| Precision| 0.27273 | 0.12500 | 0.12500 | 0.37500 | 0.12500 | 0.25000 | 0.62500 |
| Recall  | 0.42857 | 0.14286 | 0.14286 | 0.42857 | 0.14286 | 0.28571 | 0.71429 |

### Table 3. Performance of similarity measures in detecting outliers using Lenses dataset

| Measure | Overlap | Eskin | IOF | OF | Lin | Goodall4 | CPL |
|---------|---------|-------|-----|----|-----|----------|-----|
| TP      | 3       | 3     | 3   | 3  | 0   | 1        | 3   |
| TN      | 7       | 7     | 7   | 9  | 10  | 11       | 12  |
| FP      | 8       | 8     | 8   | 6  | 5   | 4        | 3   |
| FN      | 1       | 1     | 1   | 1  | 4   | 3        | 1   |
| Accuracy| 0.53000 | 0.53000 | 0.53000 | 0.63000 | 0.53000 | 0.63000 | 0.79000 |
| Error rate| 0.47000 | 0.47000 | 0.47000 | 0.37000 | 0.47000 | 0.37000 | 0.21000 |
| Precision| 0.27000 | 0.27000 | 0.27000 | 0.33000 | 0.00000 | 0.20000 | 0.50000 |
| Recall  | 0.75000 | 0.75000 | 0.75000 | 0.75000 | 0.00000 | 0.25000 | 0.75000 |

### Table 4. Performance of similarity measures in detecting outliers using Car dataset

| Measure | Overlap | Eskin | IOF | OF | Lin | Goodall4 | CPL |
|---------|---------|-------|-----|----|-----|----------|-----|
| TP      | 5       | 3     | 2   | 1  | 5   | 4        | 5   |
| TN      | 59      | 57    | 60  | 59 | 63  | 63       | 64  |
| FP      | 6       | 8     | 5   | 6  | 2   | 2        | 1   |
| FN      | 0       | 2     | 3   | 4  | 0   | 1        | 0   |
| Accuracy| 0.91000 | 0.86000 | 0.89000 | 0.86000 | 0.97000 | 0.96000 | 0.99000 |
| Error rate| 0.09000 | 0.14000 | 0.11000 | 0.14000 | 0.03000 | 0.04000 | 0.01000 |
| Precision| 0.45000 | 0.27000 | 0.29000 | 0.14000 | 0.71000 | 0.67000 | 0.83000 |
| Recall  | 1.00000 | 0.60000 | 0.40000 | 0.20000 | 1.00000 | 0.80000 | 1.00000 |

![Evaluation-Accuracy](image1)

![Evaluation-Error Rate](image2)

**Fig.2** Accuracy of various similarity measures
Correlation and Probability Based Similarity Measure for Detecting Outliers in Categorical Data

VII. CONCLUSION

In this paper we presented a novel method for determining the similarity between two data objects based on the correlation between each attribute of the data object and the class attribute, and also based on the probability of the occurrence of each attribute in the data set. The performance of the existing similarity measures were analyzed using real world datasets from the UCI machine learning repository and also compared the performance of these measures with the performance of the proposed correlation weighted probability and level based similarity measure. We used accuracy, error rate, precision and recall as metrics for evaluating the various similarity measures. We have found that our proposed CPL similarity measure outperforms all other similarity measures for different datasets. The proposed CPL similarity measure has higher values for accuracy, precision and recall compared to all other similarity measures and also the CPL measure has the least error rate.

REFERENCES
1. J. Han, M. Kamber, and J. Pei. “Data Mining: Concepts and Techniques, Massachusetts (US)”, Morgan Kaufmann, 2012.
2. Charu C. Aggarwal, "Outlier Analysis", Springer, 2013
3. S. Borah, V. Chandola, and V. Kumar, “Similarity measures for categorical data: a comparative evaluation”. In Proceedings of the eight SIAM international conference on data mining , pages 243–254, 2008.
4. L. A. Goodman, and W. H. Kruskal, “Measures of association for cross classifications”, Journal of the American Statistical Association, 49 (1954), 732-64.
5. R. H. Somers, "A Similarity Between Goodman and Kruskal’s Tau and Kendall’s Tau,with a Partial Interpretation of the Latter”. Journal of the American Statistical Association, Vol. 57, No. 300 (1962), pp. 804-812
6. D. Hawkins, "Identification of Outliers". Chapman and Hall, London and New York, 1980.
7. A. Patcha, and J.M. Park, “An overview of anomaly detection techniques: Existing solutions and latest technological trends”, Computer Networks, 2007. https://doi.org/10.1016/j.comnet.2007.02.001.
8. D. Ienco, R. G. Penza, R. Meo, "A Semisupervised Approach to the Detection and Characterization of Outliers in Categorical Data”. IEEE Transactions on Neural Networks and Learning Systems, Vol.28, No.5, 1017–1029 (2017), doi:10.1109/tnnls.2016.2526063
9. V. Chandola, A. Banerjee, V. Kumar, "Anomaly detection: A survey", ACM Comput. Surveys, vol. 41, no. 3, pp. 1–58, 2009.

AUTHORS PROFILE

First Author. Roy Thomas is a Research Scholar at Noorul Islam Centre for Higher Education, Kumaracoil, Tamil Nadu. He received his Master’s degree in Computer Science and Engineering from Anna University and is now working in Government Women’s Polytechnic College, Thiruvananthapuram as Head of Department in Computer Engineering. He has 15 years of teaching experience and his area of interest are Data mining and Machine Learning.

Second Author. Dr. J.E. Judith is a Professor in Computer Science and Engineering at Noorul Islam Centre for Higher Education, Kumaracoil, Tamil Nadu. She received her Master’s degree in Computer Science and Engineering from Karunya University and Ph.D from Noorul Islam University. She has 14 years of teaching and research experience and has published many research papers in National and International journals. Her area of interest are Data mining, Hadoop and Distributed systems.