2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

# Coupled Fuzzy $k$-Nearest Neighbors Classification of Imbalanced Non-IID Categorical Data

Chunming Liu[1], Longbing Cao[1] and Philip S Yu[2]
[1]Advanced Analytics Institute, University of Sydney Technology, Australia
Chunming.Liu@student.uts.edu.au, LongBing.Cao@uts.edu.au
[2]Computer Science, University of Illinois at Chicago, USA
psyu@cs.uic.edu

***Abstract*— Mining imbalanced data has recently received increasing attention due to its challenge and wide applications in the real world. Most of the existing work focuses on numerical data by manipulating the data structure which essentially changes the data characteristics or developing new distance or similarity measures which are designed for data with the so-called IID assumption, namely data is independent and identically distributed. This is not consistent with the real-life data and business needs, which request to fully respect the data structure and coupling relationships embedded in data objects, features and feature values. In this paper, we propose a novel coupled fuzzy similarity-based classification approach to cater for the difference between classes by a fuzzy membership and the couplings by coupled object similarity, and incorporate them into the most popular classifier: $k$NN to form a coupled fuzzy $k$NN (ie. CF-$k$NN). We test the approach on 14 categorical data sets compared to several $k$NN variants and classic classifiers including C4.5 and NaiveBayes. The experimental results show that CF-$k$NN outperforms the baselines, and those classifiers incorporated with the proposed coupled fuzzy similarity perform better than their original editions.**

## I. Introduction

Classification [1] is a widely accepted machine learning and data mining technique of great practical importance. By building appropriate classifiers, it identifies which class a new instance belongs to, based on training instances whose category memberships are known. The majority of classic classification algorithms, e.g. $k$NN, Decision Tree, Bayesian Networks and SVM [2], has been built for class-balanced data sets, i.e., each class of the data set includes a comparable number of instances. By contrast, the classification analysis on the class-imbalanced datasets (i.e. the number of instances in one class is dramatically different from that of the other one) has received much less attention, especially for the categorical data described by categorical features. It has been observed that such algorithms do not perform as good on imbalanced datasets as on balanced datasets. Hence, classifying class-imbalanced data emerges and attracts increasing attention in recent years.

In general, existing class-imbalanced classification methods represent two sorts of efforts, either manipulating the data distribution by over or under sampling or modifying existing methods to fit class imbalance. Although sampling-based methods show to outperform the original algorithms in most situation, they do not introduce much improvement for $k$NN, especially on imbalanced categorical data. This may be partially explained by the maximum-specificity induction bias of $k$NN in which the classification decision is made by examining the local neighbourhood of query instances, and therefore the global re-sampling strategies may not have pronounced effect in the local neighbourhood under examination. In addition, sampling strategies inevitably change the inherent structures of the original data, or even worse, lose information or add noise. Instead, several distance or similarity-based classification algorithms are proposed, such as $k$ENN[3] and CCW-$k$NN[4], to adapt $k$NN to imbalanced data. However, they were designed for numeric data.

Let us take some of the UCI Breast Cancer data (Table I) as an example to illustrate the problems with the existing algorithms and show the challenge of classifying class-imbalanced categorical data. As shown in Table I, eleven instances are divided into two classes with four categorical features: age, tumor-size, inv-nodes and breast-quad. The value in the brackets indicates the frequency of the corresponding feature value. It is a class-imbalanced categorical data set, since there are only three instances in class $A$ while eight instances in class $B$. Here, we use the first instance $\{u_0\}$ as the testing data set, and the rest $\{u_i\}_{i=1}^{10}$ as the training data set. If we use the traditional $k$NN algorithm to classify $u_0$, it will be labeled as $B$ due to a relatively large number of the instances in class $B$. As indicated in Table I, the Overlap Similarity is defined as

$$\text{Sim\_Overlap}(u_i, u_j) = \frac{|u_i \bigcap u_j|}{\min\{|u_i|, |u_j|\}}, \tag{1}$$

the Overlap Similarity between $(u_0, u_1)$ is equal to that of $(u_0, u_4)$, $(u_0, u_6)$, $(u_0, u_9)$ and $(u_0, u_{10})$, all are 0.5, while less than $Sim\_Overlap(u_0, u_7)$, which is the max value - 0.75. If we adopt the Cosine Similarity which is defined as

$$\text{Sim\_Cosine}(u_i, u_j) = \frac{u_i \cdot u_j}{||u_i||\,||u_j||}, \tag{2}$$

then the instances $u_{10}$, $u_1$, $u_6$ and $u_7$ will be the top 4 instances which are close to $u_0$, while $u_2$ is only the seventh close instance to $u_0$. Under this scenario, $u_0$ will be assigned to class $B$ rather than class $A$ no matter what $k$ we choose in $k$NN, because there are always more nearest neighbors labeled as class $B$ than as class $A$. Therefore, $k$NN fails to correctly classify $u_0$ within the class-imbalanced categorical data shown in Table I.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

TABLE I
AN EXAMPLE FROM THE UCI DATASET: BREAST CANCER DATA

| ID | age | tumor-size | inv-nodes | breast-quad | Class | Overlap Similarity | Cosine Similarity |
|---|---|---|---|---|---|---|---|
| $u_0$ | 50-59 | 35-39 | 0-2 | left low | A | | |
| $u_1$ | 50-59 (6) | 25-29 (2) | 0-2 (8) | right up (1) | A | 0.5 | 0.9905 |
| $u_2$ | 60-69 (1) | 30-34 (2) | 0-2 (8) | central (2) | A | 0.25 | 0.8681 |
| $u_3$ | 40-49 (2) | 25-29 (2) | 0-2 (8) | left up (4) | B | 0.25 | 0.8947 |
| $u_4$ | 50-59 (6) | 30-34 (2) | 6-8 (1) | left low (2) | B | 0.5 | 0.7274 |
| $u_5$ | 30-39 (1) | 10-14 (3) | 0-2 (8) | right low (1) | B | 0.25 | 0.8452 |
| $u_6$ | 50-59 (6) | 50-54 (1) | 0-2 (8) | left up (4) | B | 0.5 | 0.9834 |
| $u_7$ | 50-59 (6) | 35-39 (1) | 0-2 (8) | left up (4) | B | 0.75 | 0.9834 |
| $u_8$ | 50-59 (6) | 10-14 (3) | 3-5 (1) | left up (4) | B | 0.25 | 0.6817 |
| $u_9$ | 40-49 (2) | 10-14 (3) | 0-2 (8) | left low (2) | B | 0.5 | 0.9000 |
| $u_{10}$ | 50-59 (6) | 15-19 (1) | 0-2 (8) | central (2) | B | 0.5 | 1.0000 |

Besides the class imbalance, another key complexity which hasn't been catered for in existing classification algorithms like $k$NN is the comprehensive coupling relationships between feature values, features and between instances hidden in data while computing the similarity/distance between instances. Considering such couplings has shown [5] to be very important for capturing the non-IIDness nature in the real-world data, in which objects and object properties are coupled and personalized rather than independent and identically distributed as we usually assume. This is particularly important for big data analytics of complex behavioral and social data with diverse interactions.

Incorporating the couplings into classifiers relies on defining new similarity metrics which can capture the interactions between values, features and objects. This is much more doable for numerical data than categorical one, since the existing metrics such as Manhattan and euclidean distance and coefficient were mainly built for numeric variables. Matching [6] is the most common way to measure the similarity of categorical data. The overlap similarity between two categorical values is to assign 1 if they are identical otherwise 0 if different. Further, for two multivariate categorical data points, the similarity between them will be proportional to the number of features in which they match. This will cause problems in some situations. For example, considering a categorical data set $D$, which has only two features: color and size. Color takes three possible values: red, green, blue, and size takes three values: small, medium and large. Table II shows the frequency of co-occurrences of the two features.

Based on the feature values given by data set $D$, the overlap similarity between the two instances (green, small) and (green, medium) is $\frac{1}{2}$, and the overlap similarity between (blue, small) and (blue, medium) is also $\frac{1}{2}$ . But the frequency distribution in Table II shows that (green, small) and (green, medium) are frequent co-occurrences, while (blue, small) and (blue, medium) are very rare co-occurrences. Hence, the overlap measure is too simplistic by just giving the equal importance to matches and mismatches, and the co-occurrence information in categorical data reflects the interaction between features and can be useful to define what makes two categorical values more or less similar. However, such co-occurrence information hasn't been incorporated into the existing similarity metrics including the cosine similarity.

The above analysis shows that it is much challenging but essential to classify class-imbalanced non-IID categorical data. In fact, learning from the class-imbalanced data has also been identified as one of the top 10 challenging problems in data mining research [7]. To the best of our knowledge, no existing research is available for handling this, and it is not possible for the existing classification algorithms built for class-balanced IID data to capture the coupling relationships between imbalanced classes and between categorical features in the increasingly seen social networks, social media, recommender systems and behavioral applications.

In this paper, we propose a novel coupled fuzzy nearest neighbor classification algorithm, CF-$k$NN for short, for class-imbalanced non-IID categorical data. CF-$k$NN advances the idea of classic $k$NN substantially to address both class imbalance and couplings within data in terms of the following main mechanism:

- By incorporating the fuzzy set theory, CF-$k$NN assigns the corresponding size memberships to distinct classes according to their sizes to handle the multi-classes scenario in a fuzzy way.
- By exploring the feature's weight, CF-$k$NN can extract the inner coupled relationship between features and classes.
- CF-$k$NN captures the *intra-feature couplings* namely interactions within each categorical feature and *inter-feature couplings* between different categorical features and produces a similarity metric that can extract the similarity hidden at different levels in categorical data, from feature values to features and instances.

Tested on different real-life data sets from UCI[8], KEEL[9] and even an university database, CF-$k$NN outperforms the typical $k$NN algorithms, including classic $k$NN, $k$ENN which finds exemplar training samples to enlarge the decision boundary for the minority class, CCW-$k$NN which learns the class weight for each training sample by mixture modelling, and SMOTE based $k$NN which uses SMOTE to pre-process the dataset, showing its significant advantage in handling class-imbalanced data by considering couplings. The improved performance of variants $k$NN algorithms which use our coupled fuzzy strategy indicates that CF-$k$NN can better capture the intrinsic interactions and imbalance.

The paper is organized as follows. Section II briefly re-

views the related work. Preliminary definitions are specified in Section III. Section IV explains our classification algorithm on the class-imbalanced data sets. The experimental results are discussed in Section V. The conclusion and future work are summarized in Section VI.

## II. Related Work

In recent years, different approaches have been proposed to handle class imbalance classification problems. In summary, these methods can be broadly grouped into three different ways: data sampling, algorithmic modification and ensemble approach. The data sampling-based methods intend to balance the data. The common strategies are to reduce the majority class samples (undersampling) or to add new minority class samples (oversampling)[10], [11]. For instance, SMOTE [10] over-samples the minority class by taking each minority class sample and introducing synthetic examples along the line segments joining all of the k minority class nearest neighbors based on the nearest neighbor analogy. It beats the random over-sampling by adding new instances to a minority class, without suffering from the over-fitting. These methods are designed more suitable for numerical data sets. SMOTE would introduce noisy points if it is used for categorical data. Unlike sampling-based methods which change the original data structure, the approaches of modifying existing algorithms aim to make them more effective in dealing with imbalanced data, while keeping the data structure unchanged. For example, CCPDT[12], which is designed for imbalanced situation, is a modification of the decision tree algorithm. The ensemble approach incorporate approaches at the data level and algorithmic level, considering higher costs for the misclassification of examples of the positive class with respect to the negative class, and trying to minimize higher cost errors[13].

Although $k$NN has been identified as one of the top ten most influential data mining algorithms [2], the classic $k$NN algorithm is not suitable for the presence of imbalanced class distribution. To improve the performance of $k$NN for imbalanced classification, $k$ENN[3] and CCW-$k$NN[4] have been proposed. $k$ENN proposed a training stage where exemplar positive training instances are identified and generalized into Gaussian balls as concepts for the minority class. When classifying a query instance using its k nearest neighbors, the positive concepts formulated at the training stage ensure that classification is more sensitive to the minority class. This approach is based on extending the decision boundary for the minority class. CCW-$k$NN uses the probability of attribute values given class labels to weight prototypes in $k$NN. They used conditional probabilities of classes but not the probabilities of class labels in the neighborhood of the query instance. These methods perform more accurately than $k$NN. However, both $k$ENN and CCW-$k$NN were for numerical data, and require a training stage either to find exemplar training samples to enlarge the decision boundaries for the positive class, or to learn the class weight for each training sample by mixture modelling and Bayesian network learning.

TABLE II
Frequency of feature co-occurrences

| | ***small*** | ***medium*** | ***large*** | **Total** |
|---|---|---|---|---|
| ***red*** | 44 | 47 | 9 | 100 |
| ***green*** | 48 | 45 | 7 | 100 |
| ***blue*** | 8 | 8 | 84 | 100 |
| **Total** | 100 | 100 | 100 | |

Yang Song et al. [14] propose two new $k$NN algorithms based on the concept: informativeness, which is introduced as a query-based distance metric. A data point is treated informative if it is close to the query point and far away from the points with different class labels. Locally Informative $k$NN(LI-$k$NN) applies this to select the most informative points and predict the label of a query point based on the most numerous class with the neighbors; Globally Informative $k$NN(GI-$k$NN) finds the globally informative points by learning a weight vector from the training points.

The above work introduces new learning algorithms to deal with the imbalanced class distribution problem mainly for numerical data. The overlap similarity or cosine similarity[15] for categorical data is too vague to clearly describe how close two categorical instances are. Those similarity measures assume that the categorical features are independent to each other and thus the data is IID. However, with the appearance of big data application, an increasing number of researchers argue that the similarity between categorical feature values is also dependent on the couplings with other features [16], as features are more or less coupled. This brings the critical issue of learning from non-IID data, a very rarely explored topic in the data mining community.

Classifying non-IID categorical data is challenging, which needs to consider the explicit and implicit couplings between objects, features and feature values. This hasn't been paid much attention in the existing classification methods, as they were mainly designed with the IIDness assumption. Very recently, Wang et al. [5] present a coupled nominal similarity to examine both the intra-coupling and inter-coupling of categorical features. Their approaches were used for clustering class-balanced categorical data. Inspired by their work, this paper incorporates the couplings between objects, features and feature values into the classification of class-imbalanced categorical data, which has not been addressed so far.

## III. Preliminary

Classification of the class-imbalanced categorical data can be formally described as follows: $U = \{u_1, \cdots, u_m\}$ is a set of $m$ instances; $F = \{a_1, \cdots, a_n\}$ is a set of $n$ categorical features; $C = \{c_1, \cdots, c_L\}$ is a set of $L$ classes, in which each class has dramatically different numbers of instances. The goal is to classify an unlabeled testing instance $u_t$ based on the instances in the training set $\{u_i\}$ with known classes. For example, Table I exhibits a class-imbalanced data set. The training set consists of ten objects $\{u_1, u_2, \cdots, u_{10}\}$, four categorical features $\{age, tumor-size, inv-nodes, breast-quad\}$, and two classes $\{A, B\}$. There are only two instances in class $A$, while eight instances

in class $B$. Our task is to find a suitable classification model to categorize $u_0$ into class $A$.

In the following sections, the *size* of a class refers to the number of instances in this class. When we say a class $c_l$ is smaller (or larger) than $c_k$, it means that the size of class $c_l$ is smaller (or larger) than that of $c_k$. A minority class has a relatively small size, while a majority class has a relatively large size. In addition, $|H|$ is the number of instances in set $H$.

## IV. Coupled Fuzzy $k$NN

In this section, a coupled fuzzy $k$NN algorithm (i.e. CF-$k$NN for short) is proposed to handle the similarity-based classification problem on the class-imbalanced categorical data sets.

Compared to classic $k$NN, CF-$k$NN consists of three new components: *membership assignment*, *similarity calculation*, and *integration*. At the phase of membership assignment, we introduce a fuzzy membership to handle the class-imbalanced issue: *Sized Membership of Class*. This membership provides the quantification on how small a class is. Simultaneously, at the step of similarity calculation, we introduce the *Adapted Coupled Nominal Similarity* following the idea in [5] to describe the closeness between two different instances by considering both intra and inter-feature couplings and their combination. Finally, at the stage of integration, we propose the *Integrated Similarity* to measure the similarity between the test instance and the training instance by merging the adapted coupled nominal similarity and fuzzy membership of a class. The classification result of a test instance is determined according to the integrated pairwise similarity. Below, we specify these building blocks one by one.

### A. Membership Assignment

In this part, we propose a membership: *Sized Membership of Class* to characterize the structure of imbalanced classes and to capture the prior knowledge integrated from the instances.

*1) Sized Membership of Class:* In a class-imbalanced data set, there are usually several small classes that contain much less instances (i.e. minority), while a lot more instances are in some large classes (i.e. majority). However, what exactly does a small class mean? How do we quantify a small class? As it would be too reductive to regard the smallest class as the minority, we use a fuzzy way [17] to measure how small a class is according to its size, and make our approach suitable for multi-class problems. Accordingly, we have:

*Definition 1:* The **Sized Membership of Class** $\theta(\cdot)$ denotes the rate of a class $c_l$ that belongs to the minority. Formally, $\theta(\cdot)$ is defined as:

$$\theta(c_l) = 1 - \frac{|c_l|}{m}, \tag{3}$$

where $|c_l|$ is the number of instances in classes $c_l$ and $m$ is the total number of instances in the data set. Accordingly, we have $\theta(c_l) \in (0, 1)$.

The sized membership of class describes how small a class is. In special cases, $\theta(c_l)$ reaches the maximum if $c_l$ has the smallest number of instances; $\theta(c_l)$ is down to the minimum if $c_l$ is the largest class. For other medium classes, the corresponding sized membership of class falls within $(\theta(c_l)^{min}, \theta(c_l)^{max})$. When a data set is balanced with two classes, where we have $\theta(c_l) = 0.5$. In Table I, for instance, we have $\theta(c_A) = 1 - 2/10 = 4/5$, and $\theta(c_B) = 1 - 8/10 = 1/5$.

Later in measuring the similarity of instances, we will incorporate the sized membership of class $\theta(\cdot)$ into the integrated similarity measure to balance the impact of class size in measuring instance similarity.

*2) Feature Weighting:* Less relevant features that provide little information for classification should be assigned low weights, while features that provide more reliable information should be assigned higher weights. Towards this goal, the *mutual information* (MI)[18] between the values of a feature and the class of the training examples can be used to assign feature weights. Formally, we have:

*Definition 2:* The **feature weight** describes the importance degree of each categorical feature $f_j$:

$$\alpha_j = \sum_{v \in V_f} \sum_{c_j \in C} p(c_j, x_f = v) \cdot \log \frac{p(c_j, x_f = v)}{p(c_j) \cdot p(x_f = v)} \tag{4}$$

where $p(c_j)$ is the frequency of class $c_j$ among the training set D and $p(x_f = v)$ is the frequency of value $v$ for $f$ among instances in D.

This equation assigns zero to features that provide no information about the class, and a value proportional to $\log(|C|)$ to features that completely determine the class (i.e., assuming a uniform distribution on classes). As in the example of Table I, we have the normalized feature weights: $\alpha_1 = 0.2639, \alpha_2 = 0.1528, \alpha_3 = 0.3750$, and $\alpha_4 = 0.2083$.

### B. Similarity Calculation

The similarity between instances is defined for the class-imbalanced categorical data. The usual way to deal with the similarity between two categorical instances is the cosine similarity on frequency and overlap similarity on feature category. However, they are too rough to measure the similarity and they do not consider the coupling relationships among features. Wang et al. [5] introduce a coupled nominal similarity (COS) for categorical data, which addresses both the intra-coupling similarity between values within a feature and the inter-coupling similarity among different features. The proposed similarity measure picks up both explicit and implicit interactions between objects, features and feature values, and has been shown to outperform the SMS and the ADD[19] in the clustering learning. Here, we adapt the COS in our classification algorithm as follows.

*Definition 3:* Given a training data set D, a pair of values $v_j^x, v_j^y (v_j^x \neq v_j^y)$ of feature $a_j$. $v_j^x$ and $v_j^y$ are defined to be intra-related in feature $a_j$. The **Intra-Coupled Similarity** (IaCS) between feature values $v_j^x$ and $v_j^y$ of feature $a_j$ in

either training or testing data is formalized as:

$$\delta^{Ia}(v_j^x, v_j^y) = \frac{RF(v_j^x) \cdot RF(v_j^y)}{RF(v_j^x) + RF(v_j^y) + RF(v_j^x) \cdot RF(v_j^y)}, \tag{5}$$

where $RF(v_j^x)$ and $RF(v_j^y)$ are the relative occurrence frequency of values $v_j^x$ and $v_j^y$ in feature $a_j$, respectively.

The Intra-Coupled Similarity just reflects the interaction of two values in the same feature. The higher these frequencies are, the closer such two values are. Thus, Equation (5) is designed to capture the value similarity in terms of occurrence times by taking into account the frequencies of categories. Besides, since $1 \leq RF(v_j^x), RF(v_j^y) \leq m$, then $\delta^{Ia} \in [1/3, m/(m+2)]$. For example, in Table I, values $(leftlow)$ and $(leftup)$ of feature $breast-quad$ are observed 2 and 4 times respectively, so $\delta^{Ia}((leftlow), (leftup)) = (2*4)/(2+4+2*4) = 4/7$.

In contrast, the Inter-Coupled Similarity below is defined to capture the interaction between two values of two different features of an instance from either training or testing data.

*Definition 4:* For a training data set D and two different features $a_i$ and $a_j$ $(i \neq j)$, two feature values $v_i^x, v_j^y (i \neq j)$ from features $a_i$ and $a_j$, respectively, $v_i^x$ and $v_j^y$ are inter-related if there exists at least one pair value $(v_p^{xy})$ that co-occurs in features $a_i$ and $a_j$ of instance $U_p$. The **Inter-Coupled Similarity** (IeCS) between feature values $v_i^x$ and $v_j^y$ of features $a_i$ and $a_j$ is formalized as:

$$\delta_{i|j}^{Ie}(v_i^x, v_j^y) = \frac{F(v_p^{xy})}{\max(RF(v_i^x), RF(v_j^y))}, \tag{6}$$

where $F(v_p^{xy})$ is the co-occurrence frequency count function with value pair $v_p^{xy}$, and $RF(v_i^x)$ and $RF(v_j^y)$ is the relative occurrence frequency in their features respectively.

Accordingly, we have $\delta_{i|j}^{Ie} \in [0, 1]$. The Inter-Coupled Similarity reflects the interaction or relationship of two categorical values from two different features. In Table I, for example, as $\delta_{1|4}^{Ie}((60\text{-}69), (central)) = 1/\max(1, 2) = 0.50 > \delta_{1|4}^{Ie}((50\text{-}59), (central)) = 1/\max(6, 2) = 0.167$, so between features 1 and 4, the value pair [(60-69),(central)] is closer than the value pair [(50-59),(central)].

Though the superiority of COS has been verified for clustering [5], it cannot be directly used for classification, due to its lack of class information. To make COS adaptive to classification, we incorporate class label information into COS via the feature weighting. First, the correspondence problem in relation to mapping between the feature values and the classes needs to be solved. The optimal correspondence can be obtained by using the Hungarian method [20] with $O((n_j)^3)$ complexity for $n_j$ feature values. Below, the correspondence mapping is built for each feature $f_j$ $(1 \leq j \leq n)$ and a set of classes $C$.

By taking into account the feature importance, the *Adapted Coupled Object Similarity* between instances $u_{i_1}$ and $u_{i_2}$ is formalized as:

$$\begin{aligned} &AS(u_{i_1}, u_{i_2}) \\ &= \sum_{j=1}^{n} [\beta \cdot \alpha_j \delta_j^{Ia} + (1-\beta) \cdot \sum_{k=1, k\neq j}^{n} \delta_{j|k}^{Ie}] \\ &= \sum_{j=1}^{n} [\beta \cdot \alpha_j \delta_j^{Ia}(v_j^{i_1}, v_j^{i_2}) + (1-\beta) \cdot \sum_{k=1, k\neq j}^{n} \delta_{j|k}^{Ie}(v_j^{i_1}, v_k^{i_2})], \end{aligned} \tag{7}$$

where $\beta \in [0, 1]$ is the parameter that decides the weight of intra-coupled similarity, $v_j^{i_1}$ and $v_j^{i_2}$ are the values of feature $j$ for instances $u_{i_1}$ and $u_{i_2}$, respectively. $\delta_j^{Ia}$ and $\delta_{j|k}^{Ie}$ are the intra-coupled feature value similarity and inter-coupled feature value similarity, respectively. It is remarkable to note that $\alpha_j$ is the feature weight defined in Equation (4), rather than $\alpha_j = 1/n$ assumed in [15].

### C. Integration

Finally, we aggregate the membership assignment and similarity calculation, and propose an *Integrated Similarity* for classifying class-imbalanced categorical data. In this way, we are able to select distinct similarity measures for the instances with different prior belonging memberships and from different classes. The integrated similarity is defined below:

*Definition 5:* The **Integrated Similarity** represents the adapted coupled similarity measure by taking into account the feature weight, feature values' intra and features inter coupled relationship as well as the class size information. Formally,

$$IS(u_e, u_i) = \theta(C(u_i)) \cdot AS(u_e, u_i), \tag{8}$$

where $u_e$ and $u_i$ are the instances, respectively; $C(u_i)$ denotes the class of $u_i$; $\theta(\cdot)$ is the sized membership of class defined in Equation (3); and $AS(\cdot)$ is the adapted coupled object similarity defined in Equation (7).

As indicated by Equation (8), on one hand, although we only choose two classes in our experiments, the $\theta(\cdot)$ can capture the class size information, which is the key clue to the class imbalance, so it can extends to the classification tasks with multiple classes. On the other hand, the adapted similarity $AS(\cdot)$ includes not only the feature-class coupling information (feature weight), but it also capture the feature values' intra-coupling relationship and values from different features' inter-coupling relationship. These coupling relationship reflect the inner relationship of real world data. Therefore, the similarity in our algorithm is more reasonable than that in the existing similarity calculation related algorithms for the imbalanced categorical data.

### D. The CF-kNN Algorithm

As shown in Algorithm 1, the CF-$k$NN algorithm works as follows. Following the idea of $k$NN, after obtaining the Integrated Similarity between test instance $u_e$ and training instance $\{u_i\}$, we select the $k$ nearest neighbors in the training set that correspond to the $k$ highest Integrated Similarity

values. The largest class $c_l$ of the neighbors is the desired class for $u_e$. For example, in Table I, we have $IS(u_0, u_1) = 3.9785$, $IS(u_0, u_2) = 3.8054$ and $IS(u_0, u_7) = 3.8332$ to be the top three nearest neighbors to $u_0$, and so $u_0$ is categorized to its real class, namely class $A$ ($k = 3$).

**Algorithm 1** : Coupled Fuzzy $k$NN Algorithm

**Input:** An instance $u_t$ without label and a source dataset $D\{u_1, u_2, ..., u_N\}$

**Output:** The class label of $u_t$

1: Calculate the sized membership of class in the source dataset using the fuzzy set theory, and compute the feature weight
2: Create the similarity matrix which contains both intra- and inter-feature similarity for dataset $D$
3: Calculate the distance of $u_t$ to every instance in dataset $D$
4: Select $k$ points which are close to the instance $u_t$
5: Return the class label of those $k$ neighbors which has the maximum number of instances

## V. Experiments and Evaluation

### A. Data and Experimental Settings

As the publicly available data sets were often not designed for the non-IIDness test as in this work, we choose some commonly used UCI and KEEL data and a real world data set. Our motivation is that if an algorithm can show improvement on such data compared to the baselines, it has potential to differentiate itself from others in more complex data with strong couplings. In total, 14 data sets are taken from the UCI Data Repository [8], KEEL data set repository [9], and the real Student learning data taken from the records of an Australian university's students performance database (If a student failed both in course L and course S, he or she will be labeled as "Failure", or else be labeled as "Success"). In experiment 3, we use SMOTE on this student data set and created 50 new data sets with minority class varies from 1% to 50%. A short description of all the datasets is provided in Table III and the proportion of minority class to the total instances is shown as *Minority(%)*. These data sets have been selected as they typically have an imbalance class distribution (the lowest one is 0.98%). As some data sets have mixed type of features, such as D1, D2, D4 and D5, we conducted the CAIM discretization algorithm [21] on numerical features first so as to convert them into categorical ones.

We conducted 10-fold cross validation experiments to evaluate the performance of all the algorithms. In the experiments, we select not only variants of $k$NN, such as the classic $K$ Nearest Neighbors($k$NN)[2], $k$ENN[3], CCW-$k$NN[4] and SMOTE based $k$NN to compare with, but also the very popular classifiers C4.5 and NaiveBayes. To make algorithms more comparable, we further incorporate our coupled fuzzy method into some $k$NN algorithms (the new ones are with a prefix of "$CF+$") to compare their results. In all our experiments, we set $k = 5$ to all those $k$NN-based classifiers, and the confidence levels for $k$ENN is set to 0.1.

Due to the dominative effect of the majority class, the overall accuracy is not an appropriate evaluation measure for the performance of classifiers on imbalanced datasets, we use Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC)[22] to evaluate the performance results. AUC indicates the overall classification performance, and the AUC of a perfect classifier equals to 1, a bad one less than 0.5, so a good classification algorithm will has a higher AUC.

### B. The performance of CF-$k$NN

Table IV shows the AUC results for our CF-$k$NN compared with the state of the art algorithms. The top two results are highlighted in bold. Compared with other approaches, our CF-$k$NN has the highest AUC result and outperforms others in most of the datasets, especially in datasets with high imbalance rate. Also, our proposed CF-$k$NN always outperforms classic $k$NN on all the datasets. This evidences that considering the coupling relationships between objects, features and feature values by treating the data as non-IID in computing similarity or distance captures the intrinsic data characteristics. Note that the SMOTE-based $k$NN does not always demonstrate significant improvement compared with $k$NN, sometimes even worse, such as in data set D5 and D14. It means that only using SMOTE on imbalanced categorical data may not bring much improvement, but even some noise.

From the results we can see that when the imbalance rates are less than 8%, our method achieves a much better improvement (the least one is 2.08% and the highest one is 12.09%) on these very simple UCI data which does not incorporate much non-IIDness characteristics. On some specific datasets, such as D8, our methods also approach as good as CCW$k$NN. That confirms again that our coupled fuzzy strategy is very effective for imbalanced non-IID classification tasks.

### C. The effect of incorporating fuzzy membership and couplings

This set of experiments aims to test the effect of incorporating fuzzy membership of class and the coupled similarity into other classification algorithms. For doing this, we create three comparison sets by integrating the proposed coupled fuzzy mechanism into $k$ENN to form CF+$k$ENN, CCW$k$NN to form CF+CCW$k$NN, and SMOTE based $k$NN to form CF+SMOTE based $k$NN, and compare their performance. All comparable algorithms are with the same parameter settings.

Table V shows the performance results of these comparable algorithms with vs. without the coupled fuzzy mechanism. It shows that incorporating our new similarity metrics will bring more or less improvement for the classic algorithms, especially for those distance or similarity-based algorithms. This further shows that our proposed idea of incorporating the fuzzy membership of classes size and measuring the couplings between objects, features and feature values capture

TABLE III

DATA SETS, ORDERED IN THE DECREASING LEVEL OF IMBALANCE

| **Index** | **Dataset** | **Source** | **#Instances** | **#Attribute** | **#Class** | **Minority Name** | **Minority(%)** |
|---|---|---|---|---|---|---|---|
| D1 | Students | REAL | 50000 | 32 | 2 | Failure | 0.98% |
| D2 | kr-vs-k | KEEL | 28056 | 6 | 18 | five | 1.68% |
| D3 | Abalone | UCI | 4177 | 8 | 29 | Class15 | 2.47% |
| D4 | Nursery | UCI | 12960 | 8 | 5 | very recom | 2.53% |
| D5 | Dermatology | UCI | 366 | 34 | 6 | P.R.P. | 5.46% |
| D6 | Zoo | UCI | 101 | 17 | 7 | Set6 | 7.92% |
| D7 | Solar Flare | KEEL | 1066 | 11 | 6 | E | 8.91% |
| D8 | Connect-4 | UCI | 67557 | 42 | 3 | draw | 9.55% |
| D9 | Primary Tumor | UCI | 339 | 17 | 22 | stomach | 11.50% |
| D10 | Soybean(Large) | UCI | 307 | 35 | 19 | brown-spot | 13.03% |
| D11 | Hayes-roth | UCI | 160 | 5 | 3 | 3 | 19.38% |
| D12 | Contraceptive | UCI | 1473 | 9 | 3 | Long-term | 22.61% |
| D13 | Adult | UCI | 45222 | 14 | 2 | >50K | 23.93% |
| D14 | Splice-junction | KEEL | 3190 | 60 | 3 | EI | 24.04% |

TABLE IV

THE AUC RESULTS FOR CF-$k$NN IN COMPARISON WITH OTHER ALGORITHMS

| Dataset | Minority(%) | **CF-$k$NN** | $k$NN | $k$ENN | CCW$k$NN | SMOTE | C4.5 | Naive | **improvement** |
|---|---|---|---|---|---|---|---|---|---|
| D1 | 0.98% | **0.909** | 0.845 | 0.849 | 0.854 | **0.866** | 0.857 | 0.857 | 4.97%-7.59% |
| D2 | 1.68% | **0.711** | 0.661 | 0.672 | **0.685** | 0.682 | 0.669 | 0.669 | 3.87%-7.49% |
| D3 | 2.47% | **0.718** | 0.672 | 0.680 | **0.692** | 0.688 | 0.683 | 0.682 | 3.75%-6.89% |
| D4 | 2.53% | **0.981** | 0.922 | **0.959** | 0.948 | 0.933 | 0.958 | 0.934 | 2.35%-6.38% |
| D5 | 5.46% | **0.76** | 0.715 | 0.720 | **0.729** | 0.678 | 0.716 | 0.724 | 4.28%-12.09% |
| D6 | 7.92% | **0.887** | 0.842 | **0.869** | **0.869** | 0.854 | 0.857 | 0.859 | 2.08%-5.30% |
| D7 | 8.91% | **0.962** | 0.910 | 0.920 | 0.937 | 0.930 | **0.947** | 0.925 | 1.62%-5.67% |
| D8 | 9.55% | **0.916** | 0.864 | 0.876 | **0.916** | 0.910 | 0.910 | 0.888 | 0.00%-6.02% |
| D9 | 11.50% | **0.716** | 0.685 | 0.701 | 0.695 | 0.701 | 0.698 | **0.705** | 1.60%-4.59% |
| D10 | 13.03% | **0.971** | 0.932 | 0.957 | 0.961 | **0.961** | 0.942 | 0.954 | 1.01%-4.16% |
| D11 | 19.38% | **0.972** | 0.932 | 0.943 | **0.960** | 0.942 | 0.959 | 0.952 | 1.26%-4.34% |
| D12 | 22.61% | **0.755** | 0.718 | 0.729 | 0.725 | **0.743** | 0.726 | 0.736 | 1.64%-5.12% |
| D13 | 23.93% | **0.938** | 0.904 | 0.915 | 0.910 | 0.910 | **0.920** | 0.919 | 1.95%-3.79% |
| D14 | 24.04% | **0.977** | 0.938 | 0.940 | 0.947 | 0.907 | **0.964** | 0.953 | 1.36%-7.72% |

TABLE V

THE AUC RESULT COMPARISON FOR ALGORITHMS WITH AND WITHOUT COUPLED FUZZY METHOD

| Dataset | Minority(%) | $k$ENN | CF+$k$ENN | CCW$k$NN | CF+CCW$k$NN | SMOTE | CF+SMOTE |
|---|---|---|---|---|---|---|---|
| D1 | 0.98% | 0.849 | **0.905** | 0.854 | **0.906** | 0.866 | **0.922** |
| D2 | 1.68% | 0.672 | **0.715** | 0.685 | **0.726** | 0.682 | **0.725** |
| D3 | 2.47% | 0.680 | **0.724** | 0.692 | **0.733** | 0.688 | **0.735** |
| D4 | 2.53% | 0.959 | **0.979** | 0.948 | **0.967** | 0.933 | **0.990** |
| D5 | 5.46% | 0.720 | **0.766** | 0.729 | **0.771** | 0.678 | **0.718** |
| D6 | 7.92% | 0.869 | **0.922** | 0.869 | **0.918** | 0.854 | **0.908** |
| D7 | 8.91% | 0.920 | **0.975** | 0.937 | **0.989** | 0.930 | **0.985** |
| D8 | 9.55% | 0.876 | **0.928** | 0.916 | **0.965** | 0.910 | **0.963** |
| D9 | 11.50% | 0.701 | **0.742** | 0.695 | **0.732** | 0.701 | **0.741** |
| D10 | 13.03% | **0.957** | **0.957** | 0.961 | **0.973** | 0.961 | **0.975** |
| D11 | 19.38% | 0.943 | **0.990** | 0.960 | **0.974** | 0.942 | **0.995** |
| D12 | 22.61% | 0.729 | **0.764** | **0.725** | **0.725** | 0.743 | **0.776** |
| D13 | 23.93% | 0.915 | **0.957** | 0.910 | **0.946** | 0.910 | **0.951** |
| D14 | 24.04% | 0.940 | **0.981** | 0.947 | **0.984** | 0.907 | **0.947** |

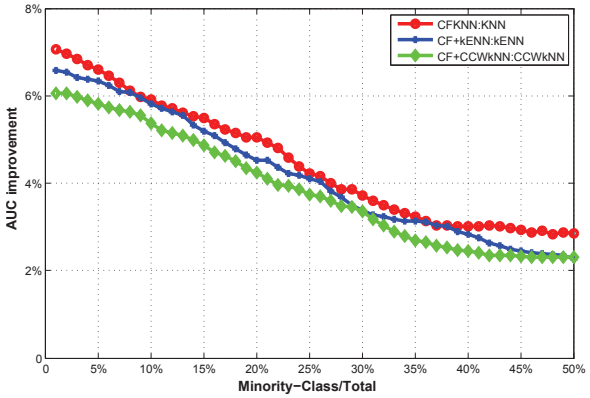


Fig. 1. The sensitivity to imbalance rate.

the intrinsic characteristics better than existing methods, and it especially suitable for class-imbalanced categorical data.

### D. The sensitivity to imbalance rate

To evaluate our coupled similarity on different imbalance rate, we do SMOTE on student data and create 50 new data sets, in which the minority class varies from 1% to 50% of the total instances. Fig. 1 shows the improvement of the basic algorithms which combined with our Coupled Fuzzy Similarity on different imbalance rate. As it shows in the figure, when minority class only takes up $< 10\%$ of the total instances, both $k$NN and $k$ENN (combined with CF) can have an improvement of over $5.821\%$. Even for CCW$k$NN, the improvement can over $5.372\%$. But with the imbalance rate declining, this improvement falls simultaneously. When minority class comes to $35\%$ of the total records (which can be defined as "balanced" data) or over, the improvement will not be so outstanding and stay stable at about $2.2\%$. This experiment demonstrates that our strategy is sensitive to the imbalance rate, and it is more suitable for being used in the scenario with high imbalance rate, that is, imbalanced categorical Non-IID data.

## VI. Conclusions and Future Work

Traditional classifiers mainly focus on dealing with balanced data set and overlook the coupling relationship between data attributes, objects and classes. Classifying coupled and imbalanced data is very challenging. We propose a coupled fuzzy $k$NN to classify imbalanced categorical data with strong relationships between objects, attributes and classes. It incorporates the size membership of a class with attribute weight into a coupled similarity measure, which effectively extracts the inter-coupling and intra-coupling relationships in categorical data. The experiment results show that our CF-$k$NN has a more stable and higher average performance than the classic $k$NN, $k$ENN, CCW$k$NN, SMOTE-based $k$NN, Decision Tree and NaiveBayes when applied on class-imbalanced categorical data. Future work will include increasing the algorithm efficiency, lowering the time complexity and extending the algorithm to mixed type data which contains both categorical features and numerical features, and even applying this idea to other basic classification algorithms based on similarity or distance, such as SVM.

## References

[1] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. Wiley-IEEE Press, 2011.
[2] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.
[3] L. Yuxuan and X. Zhang, "Improving k nearest neighbor with exemplar generalization for imbalanced classification," in *15th Pacific-Asia Conference, PAKDD 2011*. Springer, 2011, pp. 1–12.
[4] W. Liu and S. Chawla, "Class confidence weighted knn algorithms for imbalanced data sets," *Advances in Knowledge Discovery and Data Mining*, pp. 345–356, 2011.
[5] C. Wang, L. Cao, M. Wang, J. Li, W. Wei, and Y. Ou, "Coupled nominal similarity in unsupervised learning," in *CIKM 2011*. ACM, 2011, pp. 973–978.
[6] S. Santini and R. Jain, "Similarity measures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, pp. 871–883, 1999.
[7] Q. Yang and X. Wu, "10 challenging problems in data mining research," *International Journal of Information Technology & Decision Making*, vol. 5, no. 4, pp. 597–604, 2006.
[8] K. Bache and M. Lichman, "UCI machine learning repository," 2013.
[9] J. Alcalá, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2010.
[10] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.
[11] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.
[12] W. Liu, S. Chawla, D. Cieslak, and N. Chawla, "A robust decision tree algorithm for imbalanced data sets," in *SDM 2010*, 2010, pp. 766–777.
[13] R. Polikar, "Ensemble learning," in *Ensemble Machine Learning*. Springer, 2012, pp. 1–34.
[14] Y. Song, J. Huang, D. Zhou, H. Zha, and C. L. Giles, "Iknn: Informative k-nearest neighbor pattern classification," in *Knowledge Discovery in Databases: PKDD 2007*. Springer, 2007, pp. 248–264.
[15] T. Yang, L. Cao, and C. Zhang, "A novel prototype reduction method for the k-nearest neighbor algorithm with $k \geq 1$," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2010, pp. 89–100.
[16] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," in *SDM 2008*, 2008, pp. 243–254.
[17] T. J. Ross, *Fuzzy Logic with Engineering Applications*. John Wiley & Sons, 2009.
[18] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.
[19] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data and Knowledge Engineering*, vol. 63, no. 2, pp. 503–527, 2007.
[20] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
[21] L. A. Kurgan and K. J. Cios, "CAIM discretization algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 2, pp. 145–153, 2004.
[22] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.