Influence in Classification via Cooperative Game Theory

Amit Datta and **Anupam Datta** and **Ariel D. Procaccia** and **Yair Zick**
Carnegie Mellon University
amitdatta,danupam@cmu.edu
arielpro,yairzick@cs.cmu.edu

Abstract

A dataset has been classified by some unknown classifier into two types of points. What were the most important factors in determining the classification outcome? In this work, we employ an axiomatic approach in order to uniquely characterize an influence measure: a function that, given a set of classified points, outputs a value for each feature corresponding to its influence in determining the classification outcome. We show that our influence measure takes on an intuitive form when the unknown classifier is linear. Finally, we employ our influence measure in order to analyze the effects of user profiling on Google's online display advertising.

1 Introduction

A recent white house report [Podesta *et al.*, 2014] highlights some of the major risks in the ubiquitous use of big data technologies. According to the report, one of the major issues with large scale data collection and analysis is a glaring lack of transparency. For example, a credit reporting company collects consumer data from third parties, and uses machine learning analysis to estimate individuals' credit score. On the one hand, this method is "impartial": an emotionless algorithm cannot be accused of being malicious (discriminatory behavior is not hard-coded). However, it is hardly transparent; indeed, it is difficult to tease out the determinants of one's credit score: it depends on the user's financial activities, age, address, the behavior of similar users and many other factors. This is a major issue: big-data analysis does not intend to discriminate, but inadvertent discrimination does occur: treating users differently based on unfair criteria (e.g. online retailers offering different discounts or goods based on place of residence or past purchases).

In summary, big data analysis leaves users vulnerable. They may be discriminated against, and no one (including the algorithm's developers!) may even know why; what's worse, traditional methods for preserving user anonymity (e.g. by "opting out" of data collection) offer little protection; big data techniques allow companies to infer individuals' data based on similar users [Barocas and Nissenbaum, 2014]. Since it is often difficult to "pop the hood" and understand the inner workings of classification algorithms, maintaining transparency in classification is a major challenge. In more concrete terms, transparency can be interpreted as understanding what influences the decisions of a black-box classifier. This is where our work comes in.

Suppose that we are given a dataset B of users; here, every user $\mathbf{a} \in B$ can be thought of as a vector of features (e.g. $\mathbf{a} = (\text{age, gender, IP address} \dots)$), where the i-th coordinate of $\mathbf{a}$ corresponds to the state of the i-th feature. Each $\mathbf{a}$ has a value $v(\mathbf{a})$ (say, the credit score of $\mathbf{a}$). We are interested in the following question: *given a dataset B of various feature vectors and their values, how influential was each feature in determining these values?*

In more detail, given a set $N = \{1, \dots, n\}$ of features, a dataset B of feature profiles, where every profile $\mathbf{a}$ has a value $v(\mathbf{a})$, we would like to compute a measure $\phi_i(N, B, v)$ that corresponds to feature i's importance in determining the labels of the points in B. We see this work as an important first step towards a concrete methodology for transparency analysis of big-data algorithms.

Our Contribution: We take an axiomatic approach — which draws heavily on cooperative game theory — to define an influence measure. The merit of our approach lies in its independence of the underlying structure of the classification function; all we need is to collect data on its behavior.

We show that our influence measure is the unique measure satisfying some natural properties (Section 2). As a case study, we show that when the input values are given by a linear classifier, our influence measure has an intuitive geometric interpretation (Section 3). Finally, we show that our axioms can be extended in order to obtain other influence measures (Section 4). For example, our axioms can be used to obtain a measure of *state influence*, as well as influence measures where a prior distribution on the data is assumed, or a measure that uses pseudo-distance between user profiles to measure influence.

We complement our theoretical results with an implementation of our approach, which serves as a proof of concept (Section 5). Using our framework, we identify ads where certain user features have a significant influence on whether the ad is shown to users. Our experiments show that our influence measures behave in a desirable manner. In particular, a Span-

ish language ad — clearly biased towards Spanish speakers — demonstrated the highest influence of any feature among all ads.

A full version of this paper, which includes the full proofs, appears as a Cylab technical report (CMU-CyLab-15-001).[1]

1.1 Related Work

Axiomatic characterizations have played an important role in the design of provably fair revenue divisions [Shapley, 1953; Young, 1985; Banzhaf, 1965; Lehrer, 1988]. Indeed, one can think of the setting we describe as a generalization of cooperative games, where agents can have more than one state — in cooperative games, agents are either present or absent from a coalition. Some papers extend cooperative games to settings where agents have more than one state, and define influence measures for such settings [Chalkiadakis *et al.*, 2010; Zick *et al.*, 2014]; however, our setting is far more general.

Our definition of influence measures the ability of a feature to affect the classification outcome if changed (e.g. how often does a change in gender cause a change in the display frequency of an ad); this idea is used in the analysis of cause [Halpern and Pearl, 2005; Tian and Pearl, 2000], and responsibility [Chockler and Halpern, 2004]; our influence measure can be seen as an application of these ideas to a classification setting.

Influence measures are somewhat related to *feature selection* [Blum and Langley, 1997]. Feature selection is the problem of finding the set of features that are most relevant to the classification task, in order to improve the performance of a classifier on the data; that is, it is the problem of finding a subset of features, such that if we train a classifier using just those features, the error rate is minimized. Some of the work on feature selection employs feature ranking methods; some even use the Shapley value as a method for selecting the most important features [Cohen *et al.*, 2005]. Our work differs from feature selection both in its objectives and its methodology. Our measures can be used in order to rank features, but we are not interested in training classifiers; rather, we wish to decide which features influence the decision of an unknown classifier. That said, one can certainly employ our methodology in order to rank features in feature selection tasks.

When the classifier is linear, our influence measures take on a particularly intuitive interpretation as the aggregate volume between two hyperplanes [Marichal and Mossinghoff, 2006].

Recent years have seen tremendous progress on methods to enhance fairness in classification [Dwork *et al.*, 2012; Kamishima *et al.*, 2011], user privacy [Balebako *et al.*, 2012; Pedreschi *et al.*, 2008; Wills and Tatar, 2012] and the prevention of discrimination [Kamiran and Calders, 2009; Calders and Verwer, 2010; Luong *et al.*, 2011]. Our work can potentially inform all of these research thrusts: a classifier can be deemed fair if the influence of certain features is low; for example, high gender influence may indicate discrimination against a certain gender. In terms of privacy, if a hidden feature (i.e. one that is not part of the input to the classifier) has high influence, this indicates a possible breach of user privacy.

[1] `https://www.cylab.cmu.edu/research/techreports/2015/tr_cylab15001.html`

2 Axiomatic Characterization

We begin by briefly presenting our model. Given a set of *features* $N = \{1, \ldots, n\}$, let A_i be the set of possible *values*, or *states* that feature i can take; for example, the i-th feature could be gender, in which case $A_i = \{\text{male}, \text{female}, \text{other}\}$. We are given *partial* outputs of a function over a dataset containing feature profiles. That is, we are given a subset B of $A = \prod_{i \in N} A_i$, and a valuation $v(\mathbf{a})$ for every $\mathbf{a} \in B$. By given, we mean that we do not know the actual structure of v, but we know what values it takes over the dataset B. Formally, our input is a tuple $\mathcal{G} = \langle N, B, v \rangle$, where $v : A \to \mathbb{Q}$ is a function assigning a value of $v(\mathbf{a})$ to each data point $\mathbf{a} \in B$. We refer to $\mathcal{G}$ as the *dataset*. When $v(\mathbf{a}) \in \{0, 1\}$ for all $\mathbf{a} \in B$, v is a *binary classifier*. When $B = A$ and $|A_i| = 2$ for all $i \in N$, the dataset corresponds to a standard TU cooperative game [Chalkiadakis *et al.*, 2011] (and is a simple game if $v(\mathbf{a}) \in \{0, 1\}$).

We are interested in answering the following question: *how influential is feature i?* Our desired output is a measure $\phi_i(\mathcal{G})$ that will be associated with each feature i. The measure $\phi_i(\mathcal{G})$ should be a good metric of the importance of i in determining the values of v over B.

Our goal in this section is to show that there exists a unique influence measure that satisfies certain natural axioms. We begin by describing the axioms, starting with symmetry.

Given a dataset $\mathcal{G} = \langle N, B, v \rangle$ and a bijective mapping σ from N to itself, we define $\sigma\mathcal{G} = \langle \sigma N, \sigma B, \sigma v \rangle$ in the natural way: σN has all of the features relabeled according to σ (i.e. the index of i is now $\sigma(i)$); σB is $\{\sigma\mathbf{a} \mid \mathbf{a} \in B\}$, and $\sigma v(\sigma\mathbf{a}) = v(\mathbf{a})$ for all $\sigma\mathbf{a} \in \sigma B$. Given a bijective mapping $\tau : A_i \to A_i$ over the states of some feature $i \in N$, we define $\tau\mathcal{G} = \langle N, \tau B, \tau v \rangle$ in a similar manner.

Definition 2.1. An influence measure ϕ satisfies the *feature symmetry* property if it is invariant under relabelings of features: given a dataset $\mathcal{G} = \langle N, B, v \rangle$ and some bijection $\sigma : N \to N$, $\phi_i(\mathcal{G}) = \phi_{\sigma(i)}(\sigma\mathcal{G})$ for all $i \in N$. A influence measure ϕ satisfies the *state symmetry* property if it is invariant under relabelings of states: given a dataset $\mathcal{G} = \langle N, B, v \rangle$, some $i \in N$, and some bijection $\tau : A_i \to A_i$, $\phi_j(\mathcal{G}) = \phi_j(\tau\mathcal{G})$ for all $j \in N$. Note that it is possible that $i \neq j$. A measure satisfying both state and feature symmetry is said to satisfy the *symmetry* axiom (Sym).

Feature symmetry is a natural extension of the symmetry axiom defined for cooperative games (see e.g. [Banzhaf, 1965; Lehrer, 1988; Shapley, 1953]). However, state symmetry does not make much sense in classic cooperative games; it would translate to saying that for any set of players $S \subseteq N$ and any $j \in N$, the value of i is the same if we treat S as $S \setminus \{j\}$, and $S \setminus \{j\}$ as S. While in the context of cooperative games this is rather uninformative, we make non-trivial use of it in what follows.

We next describe a sufficient condition for a feature to have no influence: a feature should not have any influence if it does not affect the outcome in any way. Formally, a feature $i \in N$ is a *dummy* if $v(\mathbf{a}) = v(\mathbf{a}_{-i}, b)$ for all $\mathbf{a} \in B$, and all $b \in A_i$

such that $(\mathbf{a}_{-i}, b) \in B$; here $\mathbf{a}_{-i}$ is the vector $\mathbf{a}$ with the i-th coordinate omitted, and $(\mathbf{a}_{-i}, b)$ is the vector $\mathbf{a}$ with the i-th coordinate set to b.

Definition 2.2. An influence measure ϕ satisfies the *dummy* (D) property if $\phi_i(\mathcal{G}) = 0$ whenever i is a dummy in the dataset $\mathcal{G}$.

The dummy property is a standard extension of the dummy property used in value characterizations in cooperative games. However, when dealing with real datasets, it may very well be that there is no vector $\mathbf{a} \in B$ such that $(\mathbf{a}_{-i}, b) \in B$; this issue is discussed further in Section 6.

Cooperative game theory employs a notion of value additivity in the characterization of both the Shapley and Banzhaf values. Given two datasets $\mathcal{G}_1 = \langle N, B, v_1\rangle, \mathcal{G}_2 = \langle N, B, v_2\rangle$, we define $\mathcal{G} = \langle N, A, v\rangle = \mathcal{G}_1 + \mathcal{G}_2$ with $v(\mathbf{a}) = v_1(\mathbf{a}) + v_2(\mathbf{a})$ for all $\mathbf{a} \in B$.

Definition 2.3. An influence measure ϕ satisfies additivity (AD) if $\phi_i(\mathcal{G}_1 + \mathcal{G}_2) = \phi_i(\mathcal{G}_1) + \phi_i(\mathcal{G}_2)$ for any two datasets $\mathcal{G}_1 = \langle N, B, v_1\rangle, \mathcal{G}_2 = \langle N, B, v_2\rangle$.

The additivity axiom is commonly used in the axiomatic analysis of revenue division in cooperative games (see [Lehrer, 1988; Shapley, 1953]); however, it fails to capture a satisfactory notion of influence in our more general setting. We now show that any measure that satisfies additivity, in addition to the symmetry and dummy properties, must evaluate to zero for all features. To show this, we first define the following simple class of datasets.

Definition 2.4. Let $\mathcal{U}_{\mathbf{a}} = \langle N, A, u_{\mathbf{a}}\rangle$ be the dataset defined by the classifier $u_{\mathbf{a}}$, where $u_{\mathbf{a}}(\mathbf{a}') = 1$ if $\mathbf{a}' = \mathbf{a}$, and is 0 otherwise. The dataset $\mathcal{U}_{\mathbf{a}}$ is referred to as the *singleton dataset* over $\mathbf{a}$.

It is an easy exercise to show that additivity implies that for any scalar $\alpha \in \mathbb{Q}$, $\phi_i(\alpha\mathcal{G}) = \alpha\phi_i(\mathcal{G})$, where the dataset $\alpha\mathcal{G}$ has the value of every point scaled by a factor of α.

We emphasize that in all proofs appearing in this section we assume for ease of exposition that $B = A$, i.e. we are given all possible profiles. Our results can be extended to the general case (though the extension is at times non-trivial).

Proposition 2.5. *Any influence measure that satisfies the (Sym), (D) and (AD) axioms evaluates to zero for all features.*

Proof. First, we show that for any $\mathbf{a}, \mathbf{a}' \in A$ and any $b \in A_i$, it must be the case that $\phi_i(\mathcal{U}_{(\mathbf{a}_{-i},b)}) = \phi_i(\mathcal{U}_{(\mathbf{a}'_{-i},b)})$. This is true because we can define a bijective mapping from $\mathcal{U}_{(\mathbf{a}_{-i},b)}$ to $\mathcal{U}_{(\mathbf{a}'_{-i},b)}$: for every $j \in N \setminus \{i\}$, we swap a_j and a'_j. By state symmetry, $\phi_i(\mathcal{U}_{(\mathbf{a}_{-i},b)}) = \phi_i(\mathcal{U}_{(\mathbf{a}'_{-i},b)})$.

Next, if ϕ is additive, then for any dataset $\mathcal{G} = \langle N, A, v\rangle$, $\phi_i(\mathcal{G}) = \sum_{\mathbf{a}\in A} v(\mathbf{a})\phi_i(\mathcal{U}_{\mathbf{a}})$. That is, the influence of a feature must be the sum of its influenceover singleton datasets, scaled by $v(\mathbf{a})$.

Now, suppose for contradiction that there exists some singleton dataset $\mathcal{U}_{\bar{\mathbf{a}}}$ ($\bar{\mathbf{a}} \in A$) for which some feature $i \in N$ does not have an influence of zero. That is, we assume that $\phi_i(\mathcal{U}_{\bar{\mathbf{a}}}) \neq 0$. We define a dataset $\mathcal{G} = \langle N, A, v\rangle$ in the following manner: for all $\mathbf{a} \in A$ such that $\mathbf{a}_{-i} = \bar{\mathbf{a}}_{-i}$, we set $v(\mathbf{a}) = 1$, and $v(\mathbf{a}) = 0$ if $\mathbf{a}_{-i} \neq \bar{\mathbf{a}}_{-i}$. In the resulting dataset, $v(\mathbf{a})$ is solely determined by the values of features in $N \setminus \{i\}$; in other words $v(\mathbf{a}) = v(\mathbf{a}_{-i}, b)$ for all $b \in A_i$, hence feature i is a dummy. According to the dummy axiom, we must have that $\phi_i(\mathcal{G}) = 0$; however,

$$0 = \phi_i(\mathcal{G}) = \sum_{\mathbf{a}:v(\mathbf{a})=1} \phi_i(\mathcal{U}_{\mathbf{a}}) = \sum_{b\in A_i} \phi_i(\mathcal{U}_{(\bar{\mathbf{a}}_{-i},b)})$$
$$= \sum_{b\in A_i} \phi_i(\mathcal{U}_{\bar{\mathbf{a}}}) = |A_i|\phi_i(\mathcal{U}_{\bar{\mathbf{a}}}) > 0,$$

where the first equality follows from the decomposition of $\mathcal{G}$ into singleton datasets, and the third equality holds by Symmetry, a contradiction. □

As Proposition 2.5 shows, the additivity, symmetry and dummy properties do not lead to a meaningful description of influence. A reader familiar with the axiomatic characterization of the Shapley value [Shapley, 1953] will find this result rather disappointing: the classic characterizations of the Shapley and Banzhaf values assume additivity (that said, The axiomatization by Young [1985] does not assume additivity).

We now show that there is an influence measure uniquely defined by an alternative axiom, which echoes the union-intersection property described by Lehrer [1988]. In what follows, we assume that all datasets are classified by a binary classifier. We write $W(B)$ to be the set of all profiles in B such that $v(\mathbf{a}) = 1$, and $L(B)$ to be the set of all profiles in B that have a value of 0. We refer to $W(B)$ as the *winning profiles* in B, and to $L(B)$ as the *losing profiles* in B. We can thus write $\phi_i(W(B), L(B))$, rather than $\phi_i(\mathcal{G})$. Given two disjoint sets $W, L \subseteq A$, we can define the dataset as $\mathcal{G} = \langle W, L\rangle$, and the influence of i as $\phi_i(W, L)$, without explicitly writing N, B and v. As we have seen, no measure can satisfy the additivity axiom (as well as symmetry and dummy axioms) without being trivial. We now propose an alternative influence measure, captured by the following axiom:

Definition 2.6. An influence measure ϕ satisfies the *disjoint union (DU)* property if for any $Q \subseteq A$, and any disjoint $R, R' \subseteq A \setminus Q$, $\phi_i(Q, R) + \phi_i(Q, R') = \phi_i(Q, R \cup R')$, and $\phi_i(R, Q) + \phi_i(R', Q) = \phi_i(R \cup R', Q)$.

An influence measure ϕ satisfying the (DU) axiom is additive with respect to independent observations *of the same type*. Suppose that we are given the outputs of a binary classifier on two datasets: $\mathcal{G}_1 = \langle W, L_1\rangle$ and $\mathcal{G}_2 = \langle W, L_2\rangle$. The (DU) axiom states that the ability of a feature to affect the outcome on $\mathcal{G}_1$ is independent of its ability to affect the outcome in $\mathcal{G}_2$, if the winning states are the same in both datasets.

Replacing additivity with the disjoint union property yields a unique influence measure, with a rather simple form.

$$\chi_i(\mathcal{G}) = \sum_{\mathbf{a}\in B} \sum_{b\in A_i:(\mathbf{a}_{-i},b)\in B} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})| \qquad (1)$$

χ measures the number of times that a change in the state of i causes a change in the classification outcome. If we normalize χ and divide by $|B|$, the resulting measure has the following intuitive interpretation: pick a vector $\mathbf{a} \in B$ uniformly at random, and count the number of points in A_i for which $(\mathbf{a}_{-i}, b) \in B$ and i changes the value of $\mathbf{a}$. We note

that when all features have two states and $B = A$, χ coincides with the (raw) Banzhaf power index [Banzhaf, 1965].

We now show that χ is a unique measure satisfying (D), (Sym) and (DU). We begin by presenting the following lemma, which characterizes influence measures satisfying (D), (Sym) and (DU) when dataset contains only a single feature. Again, we assume here that $B = A$.

Lemma 2.7. *Let ϕ be an influence measure that satisfies state symmetry, and let $\mathcal{G}_1 = \langle\{i\}, A_i, v_1\rangle$ and $\mathcal{G}_2 = \langle\{i\}, A_i, v_2\rangle$ be two datasets with a single feature i; if the number of winning profiles under $\mathcal{G}_1$ and $\mathcal{G}_2$ is identical, then $\phi_i(\mathcal{G}_1) = \phi_i(\mathcal{G}_2)$.*

Proof Sketch. We simply construct a bijective mapping from the winning states of i under $\mathcal{G}_1$ and its winning states in $\mathcal{G}_2$. By state symmetry, $\phi_i(\mathcal{G}_1) = \phi_i(\mathcal{G}_2)$. □

Lemma 2.7 implies that for single feature games, the value of a feature only depends on the number of winning states, rather than their identity.

We are now ready to show the main theorem for this section: χ is the unique influence measure satisfying the three axioms above, up to a constant factor.

Theorem 2.8. *An influence measure ϕ satisfies (D), (Sym) and (DU) if and only if there exists a constant C such that for every dataset $\mathcal{G} = \langle N, A, v\rangle$ and all $i \in N$,*

$$\phi_i(\mathcal{G}) = C \cdot \chi_i(\mathcal{G}).$$

Proof. It is an easy exercise to verify that χ satisfies the three axioms, so we focus on the "only if" direction. Let us write $W = W(A)$ and $L = L(A)$. Given some $\mathbf{a}_{-i} \in A_{-i}$, we write $L_{\mathbf{a}_{-i}} = \{\bar{\mathbf{a}} \in L \mid \mathbf{a}_{-i} = \bar{\mathbf{a}}_{-i}\}$, and $W_{\mathbf{a}_{-i}} = \{\bar{\mathbf{a}} \in W \mid \mathbf{a}_{-i} = \bar{\mathbf{a}}_{-i}\}$.

Using the disjoint union property, we can decompose $\phi_i(W, L)$ as follows:

$$\phi_i(W, L) = \sum_{\mathbf{a}_{-i} \in A_{-i}} \sum_{\bar{\mathbf{a}}_{-i} \in A_{-i}} \phi_i(W_{\mathbf{a}_{-i}}, L_{\bar{\mathbf{a}}_{-i}}). \quad (2)$$

Now, if $\bar{\mathbf{a}}_{-i} \neq \mathbf{a}_{-i}$, then feature i is a dummy given the dataset provided. Indeed, state profiles are either in $W_{\mathbf{a}_{-i}}$ or in $L_{\bar{\mathbf{a}}_{-i}}$; that is, if $v(\mathbf{a}_{-i}, b) = 0$, then $(\mathbf{a}_{-i}, b)$ is unobserved, and if $v(\bar{\mathbf{a}}_{-i}, b) = 1$, then $(\bar{\mathbf{a}}_{-i}, b)$ is unobserved. We conclude that

$$\phi_i(W, L) = \sum_{\mathbf{a}_{-i} \in A_{-i}} \phi_i(W_{\mathbf{a}_{-i}}, L_{\mathbf{a}_{-i}}). \quad (3)$$

Let us now consider $\phi_i(W_{\mathbf{a}_{-i}}, L_{\mathbf{a}_{-i}})$. Since ϕ satisfies state symmetry, Lemma 2.7 implies that ϕ_i can only possibly depend on $\mathbf{a}_{-i}$, $|W_{\mathbf{a}_{-i}}|$ and $|L_{\mathbf{a}_{-i}}|$. Next, for any $\mathbf{a}_{-i}$ and $\mathbf{a}'_{-i}$ such that $|L_{\mathbf{a}_{-i}}| = |L_{\mathbf{a}'_{-i}}|$ and $|W_{\mathbf{a}_{-i}}| = |W_{\mathbf{a}'_{-i}}|$, so by Lemma 2.7 $\phi_i(W_{\mathbf{a}_{-i}}, L_{\mathbf{a}_{-i}}) = \phi_i(W_{\mathbf{a}'_{-i}}, L_{\mathbf{a}'_{-i}})$. In other words ϕ_i only depends on $|W_{\mathbf{a}_{-i}}|, |L_{\mathbf{a}_{-i}}|$, and not on the identity of $\mathbf{a}_{-i}$.

Thus, one can see ϕ_i for a single feature as a function of two parameters, w and l in $\mathbb{N}$, where w is the number of winning states and l is the number of losing states. According to the dummy property, we know that $\phi_i(w, 0) = \phi_i(0, l) = 0$; moreover, the disjoint union property tells us that $\phi_i(x, l) + \phi_i(y, l) = \phi_i(x + y, l)$, and that $\phi_i(w, x) + \phi_i(w, y) = \phi_i(w, x + y)$. We now show that $\phi_i(w, l) = \phi_i(1, 1)wl$.

Our proof is by induction on $w+l$. For $w+l = 2$ the claim is clear. Now, assume without loss of generality that $w > 1$ and $l \geq 1$; then we can write $w = x + y$ for $x, y \in \mathbb{N}$ such that $1 \leq x, y < w$. By our previous observation,

$$\begin{aligned}\phi_i(w, l) &= \phi_i(x, l) + \phi_i(y, l)\\ &\overset{i.h.}{=} \phi_i(1, 1)xl + \phi_i(1, 1)yl = \phi_i(1, 1)wl.\end{aligned}$$

Now, $\phi_i(1, 1)$ is the influence of feature i when there is exactly one losing state profile, and one winning state profile. We write $\phi_i(1, 1) = c_i$.

Let us write $W_i(\mathbf{a}_{-i}) = \{b \in A_i \mid v(\mathbf{a}_{-i}, b) = 1\}$ and $L_i(\mathbf{a}_{-i}) = A_i \setminus W_i(\mathbf{a}_{-i})$. Thus, $|W_{\mathbf{a}_{-i}}| = |W_i(\mathbf{a}_{-i})|$, and $|L_{\mathbf{a}_{-i}}| = |L_i(\mathbf{a}_{-i})|$. Putting it all together, we get that

$$\phi_i(\mathcal{G}) = c_i \sum_{\mathbf{a}_{-i} \in A_{-i}} |W_i(\mathbf{a}_{-i})| \cdot |L_i(\mathbf{a}_{-i})| \quad (4)$$

We just need to show that the measure given in (4) equals χ_i (modulo c_i). Indeed, (4) equals $\sum_{\mathbf{a} \in A:\ v(\mathbf{a})=0} |W_i(\mathbf{a}_{-i})|$, which in turn equals $\sum_{\mathbf{a} \in A:\ v(\mathbf{a})=0} \sum_{b \in A_i} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|$. Similarly, (4) equals

$$\sum_{\mathbf{a} \in A : v(\mathbf{a})=1} \sum_{b \in A_i} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|.$$

Thus,

$$\sum_{\mathbf{a}_{-i} \in A_{-i}} |W_i(\mathbf{a}_{-i})| \cdot |L_i(\mathbf{a}_{-i})| = \frac{1}{2} \sum_{\mathbf{a} \in A} \sum_{b \in A_i} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|;$$

in particular, for every dataset $\mathcal{G} = \langle N, A, v\rangle$ and every $i \in N$, there is some constant C_i such that $\phi_i(\mathcal{G}) = C_i \chi_i(\mathcal{G})$. To conclude the proof, we must show that $C_i = C_j$ for all $i, j \in N$. Let $\sigma : N \to N$ be the bijection that swaps i and j; then $\phi_i(\mathcal{G}) = \phi_{\sigma(i)}(\sigma\mathcal{G})$. By feature symmetry, $C_i \chi_i(\mathcal{G}) = \phi_i(\mathcal{G}) = \phi_{\sigma(i)}(\sigma\mathcal{G}) = \phi_j(\sigma\mathcal{G}) = C_j \chi_j(\sigma\mathcal{G}) = C_j \chi_i(\mathcal{G})$, thus $C_i = C_j$. □

3 Case Study: Influence for Linear Classifiers

To further ground our results, we now present their application to the class of linear classifiers. For this class of functions, our influence measure takes on an intuitive interpretation.

A *linear classifier* is defined by a hyperplane in $\mathbb{R}^n$; all points that are on one side of the hyperplane are colored blue (in our setting, have value 1), and all points on the other side are colored red (have a value of 0). Formally, we associate a weight $w_i \in \mathbb{R}$ with every one of the features in N (we assume that $w_i \neq 0$ for all $i \in N$); a point $\mathbf{x} \in \mathbb{R}^n$ is blue if $\mathbf{x} \cdot \mathbf{w} \geq q$, where $q \in \mathbb{R}$ is a given parameter. The classification function $v : \mathbb{R}^n \to \{0, 1\}$ is given by $v(\mathbf{x}) = 1$ if $\mathbf{x} \cdot \mathbf{w} \geq q$, and $v(\mathbf{x}) = 0$ otherwise.

Fixing the value of x_i to some $b \in \mathbb{R}$, let us consider the set $W_i(b) = \{\mathbf{x}_{-i} \in \mathbb{R}^{n-1} \mid v(\mathbf{x}_{-i}, b) = 1\}$; we observe that if $b < b'$ and $w_i > 0$, then $W_i(b) \subset W_i(b')$ (if $w_i < 0$ then $W_i(b') \subset W_i(b)$). Given two values $b, b' \in \mathbb{R}$, we denote by

$$D_i(b, b') = \{\mathbf{x}_{-i} \in \mathbb{R}^{n-1} \mid v(\mathbf{x}_{-i}, b) \neq v(\mathbf{x}_{-i}, b')\}.$$

By our previous observation, if $b < b'$ then $D_i(b, b') = W_i(b') \setminus W_i(b)$, and if $b > b'$ then $D_i(b, b') = W_i(b) \setminus W_i(b')$.

Suppose that rather than taking points in $\mathbb{R}^n$, we only take points in $[0, 1]^n$; then we can define $|D_i(b, b')| = Vol(D_i(b, b'))$, where

$$Vol(D_i(b, b')) = \int_{\mathbf{x}_{-i} \in [0,1]^{n-1}} |v(\mathbf{x}_{-i}, b') - v(\mathbf{x}_{-i}, b)| \partial \mathbf{x}_{-i}.$$

In other words, in order to measure the total influence of setting the state of feature i to b, we must take the total volume of $D_i(b, b')$ for all $b' \in [0, 1]$, which equals $\int_{b'=0}^{1} Vol(D_i(b, b')) \partial b$. Thus, the total influence of setting the state of i to b is $\int_{\mathbf{x} \in [0,1]^n} |v(\mathbf{x}_{-i}, b) - v(\mathbf{x})| \partial \mathbf{x}$. The total influence of i would then be naturally the total influence of its states, i.e.

$$\int_{b=0}^{1} \int_{\mathbf{x} \in [0,1]^n} |v(\mathbf{x}_{-i}, b) - v(\mathbf{x})| \partial \mathbf{x} \partial b. \tag{5}$$

The formula in Equation (5) is denoted by $\chi_i(\mathbf{w}; q)$. Equation (1) is a discretized version of Equation (5); the results of Section 2 can be extended to the continuous setting, with only minimal changes to the proofs.

We now show that the measure given in (5) agrees with the weights in some natural manner. This intuition is captured in Theorem 3.1 (proof omitted).

Theorem 3.1. *Let v be a linear classifier defined by $\mathbf{w}$ and q; then $\chi_i(\mathcal{G}) \geq \chi_j(\mathcal{G})$ if and only if $|w_i| \geq |w_j|$.*

Given Theorem 3.1, one would expect the following to hold: suppose that we are given two weight vectors, $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^n$ such that $w_j = w'_j$ for all $j \neq i$, but $w_i < w'_i$. Let v be the linear classifier defined by $\mathbf{w}$ and q and v' be the linear classifier defined by $\mathbf{w}'$ and q. Is it the case that feature i is more influential under v' than under v? In other words, does influence monotonicity hold when we increase the weight of an individual feature? The answer to this is negative.

Example 3.2. Let us consider a single feature game where $N = \{1\}$, $A_1 = [0, 1]$, and $v(x) = 1$ if $wx \geq q$, and $v(x) = 0$ if $wx < q$ for a given $w > q$. The fraction of times that 1 is pivotal is

$$|Piv_1| = \int_{b=0}^{1} \int_{x=0}^{1} \mathbb{I}(v(b)=1 \wedge v(x)=0) \partial x \partial b;$$

simplifying, this expression is equal to $\left(1 - \frac{q}{w}\right) \frac{q}{w}$. We can show that $\chi_1 = 2|Piv_1|$, we have that χ_1 is maximized when $q = 2w$; in particular, χ_1 is monotone increasing when $q < w \leq 2q$, and it is monotone decreasing when $w \geq 2q$.

Example 3.2 highlights the following phenomenon: fixing the other features to be $\mathbf{a}_{-i}$, the influence of i is maximized when $|L_{\mathbf{a}_{-i}}| = |W_{\mathbf{a}_{-i}}|$. This can be interpreted probabilistically: we sample a random feature from B, and assume that for any fixed $\mathbf{a}_{-i} \in A_{-i}$, $\Pr[v(\mathbf{a}_{-i}, b) = 1] = \frac{1}{2}$. The better a feature i agrees with our assumption, the more i is rewarded. More generally, an influence measure satisfies the *agreement with prior assumption* (APA) axiom if for any vector $(p_1, \ldots, p_n) \in [0, 1]^n$, and any fixed $\mathbf{a}_{-i} \in A_{-i}$, i's influence increases as $|\Pr[v(\mathbf{a}_{-i}, b) = 1] - p_i|$ decreases. A variant of the symmetry axiom (that reflects changes in probabilities when labels change), along with the dummy and disjoint union axioms can give us a weighted influence measure as described in Section 4.2, that also satisfies the (APA) axiom.

4 Extensions of the Feature Influence Measure

Section 2 presents an axiomatic characterization of feature influence, where the value of each feature vector is either zero or 1. We now present a few possible extensions of the measure, and the variations on the axioms that they require.

4.1 State Influence

Section 2 provided an answer to questions of the following form: what is the impact of gender on classification outcomes? The answer provided in previous sections was that influence was a function of the feature's ability to change outcomes by changing its state.

It is also useful to ask a related question: what is the impact of the gender feature being set to "female" on classification outcomes? In other words, rather than measuring feature influence, we are measuring the influence of feature i being in a certain *state*. The results described in Section 2 can be easily extended to this setting. Moreover, the impossibility result described in Proposition 2.5 no longer holds when we measure state — rather than feature — influence: we can replace the disjoint union property with additivity to obtain an alternative classification of state influence.

4.2 Weighted Influence

Suppose that in addition to the dataset B, we are given a weight function $w : B \to \mathbb{R}$. $w(\mathbf{a})$ can be thought of as the number of occurrences of the vector $\mathbf{a}$ in the dataset, the probability that $\mathbf{a}$ appears, or some intrinsic importance measure of $\mathbf{a}$. Note that in Section 2 we implicitly assume that all points occur at the same frequency (are equally likely) and are equally important. A simple extension of the disjoint union and symmetry axioms to a weighted variant shows that the only weighted influence measure that satisfies these axioms is

$$\chi_i^w(B) = \sum_{\mathbf{a} \in B} \sum_{b \in A_i : (\mathbf{a}_{-i}, b) \in B} w(\mathbf{a}) |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|.$$

4.3 General Distance Measures

Suppose that instead of a classifier $v : A \to \{0, 1\}$ we are given a pseudo-distance measure: that is, a function $d : A \times A \to \mathbb{R}$ that satisfies $d(\mathbf{a}, \mathbf{a}') = d(\mathbf{a}', \mathbf{a})$, $d(\mathbf{a}, \mathbf{a}) = 0$ and the triangle inequality. Note that it is possible that $d(\mathbf{a}, \mathbf{a}') = 0$ but $\mathbf{a} \neq \mathbf{a}'$. An axiomatic analysis in such general settings is possible, but requires more assumptions on the behavior of the influence measure. Such an axiomatic approach leads us to show that the influence measure

$$\chi_i^d(B) = \sum_{\mathbf{a} \in B} \sum_{b \in A_i : (\mathbf{a}_{-i}, b) \in B} d((\mathbf{a}_{-i}, b), \mathbf{a})$$

is uniquely defined via some natural axioms. The additional axioms are a simple extension of the disjoint union

property, and a minimal requirement stating that when $B = \{\mathbf{a}, (\mathbf{a}_{-i}, b)\}$, then the influence of a feature is $\alpha d((\mathbf{a}_{-i}, b), \mathbf{a})$ for some constant α independent of i. The extension to pseudo-distances proves to be particularly useful when we conduct empirical analysis of Google's display ads system, and the effects user metrics have on display ads.

5 Implementation

We implement our influence measure to study Google's display advertising system. Users can set demographics (like gender or age) on the Google Ad Settings page[2]; these are used by the Google ad serving algorithm to determine which ads to serve. We apply our influence measure to study how demographic settings influence the targeted ads served by Google. We use the AdFisher tool [Datta *et al.*, 2014] for automating browser activity and collect ads.

We pick the set of features: $N = \{\text{gender, age, language}\}$. Feature states are $\{male, female\}$ for gender, $\{18-24, 35-44, 55-64\}$ for age, and $\{\text{English, Spanish}\}$ for language; this gives us $2 \times 3 \times 2 = 12$ possible user profiles. Using AdFisher, we launch twelve fresh browser instances, and assign each one a random user profile. For each browser instance, the corresponding settings are applied on the Ad Settings page, and Google ads on the BBC news page `bbc.com/news` are collected. For each browser, the news page is reloaded 10 times with 5 second intervals.

To eliminate ads differing due to random chance, we collect ads over 100 iterations, each comprising of 12 browser instances, thereby obtaining data for 1200 simulated users. In order to minimize confounding factors such as location and system specifications, all browser instances were run from the same stationary Ubuntu machine. The 1200 browsers received a total of $32,451$ ads (763 unique); in order to reduce the amount of noise, we focus only on ads that were displayed more than 100 times, leaving a total of 55 unique ads. Each user profile $\mathbf{a}$ thus has a frequency vector of all ads $v'(\mathbf{a}) \in \mathbb{N}^{55}$, where the k^{th} coordinate is the number of times ad k appeared for a user profile $\mathbf{a}$. We normalize $v'(\mathbf{a})$ for each ad by the total number of times that ad appeared. Thus we obtain the final value-vectors by computing $v_k(\mathbf{a}) = \frac{v'_k(\mathbf{a})}{\sum_{\mathbf{a}} v'_k(\mathbf{a})}, \forall \mathbf{a}, \forall k \in \{1, \ldots, 55\}$.

Since user profile values are vectors, we use the general distance influence measure described in Section 4.3. The pseudo-distance we use is Cosine similarity: $cosd(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}|| \cdot ||\mathbf{y}||}$; this has been used Cosine similarity has been used by Tschantz *et al.* [2014] and Guha *et al.* [2010] to measure similarity between display ads. The influence measure for gender, age, and language were 0.124, 0.120, and 0.141 respectively; in other words, no specific feature has a strong influence over ads displayed.

We next turn to measuring feature effects on specific ads. Fixing an ad k, we define the value of a feature vector to be the number of times that ad k was displayed for users with that feature vector, and use χ to measure influence.

We compare the influence measures for each attribute across all the ads and identify the top ads that demonstrate high influence. The ad for which language had the highest influence (0.167) was a Spanish language ad, which was served only to browsers that set 'Spanish' as their language on the Ad Settings page. Comparing with statistics like mean and maximum over measures across all features given in Table 1, we can see that this influence was indeed high.

Statistic	Gender	Age	Language
Max	0.07	0.0663	0.167
Min	0.00683	0.00551	0.00723
Mean	0.0324	0.0318	0.0330
Median	0.0299	0.0310	0.0291
StdDev	0.0161	0.0144	0.024

Table 1: Statistics over influence measures across features.

To conclude, using a general distance measure between two value-vectors, we identify that language has the highest influence on ads. By using a more fine-grained distance function, we can single out one ad which demonstrates high influence for language. While in this case the bias is acceptable, the experiment suggests that our framework is effective in pinpointing biased or discriminatory ads.

6 Conclusions and Future Work

In this work, we analyze influence measures for classification tasks. Our influence measure is uniquely defined by a set of natural axioms, and is easily extended to other settings. The main advantage of our approach is the minimal knowledge we have of the classification algorithm. We show the applicability of our measure by analyzing the effects of user features on Google's display ads, despite having no knowledge of Google's (potentially complex) classification algorithm.

Dataset classification is a useful application of our methods; however, our work applies to extensions of TU cooperative games where agents have more than two states (e.g. OCF games [Chalkiadakis *et al.*, 2010]).

The measure χ is trivially hard to compute exactly, as it generalizes the Banzhaf index, for which this task is known to be hard [Chalkiadakis *et al.*, 2011]. That said, both the Shapley and Banzhaf values can be approximated via random sampling [Bachrach *et al.*, 2010]. It is straightforward to show that random sampling provides good approximations for χ as well, assuming binary classifiers.

Our results can be extended in several ways. The measure χ is the number of times a change in a feature's state causes a change in the outcome. However, a partial dataset of observations may not contain any pair of vectors $\mathbf{a}, \mathbf{a}' \in B$, such that $\mathbf{a}' = (\mathbf{a}_{-i}, b)$. In Section 5, we control the dataset, so we ensure that all feature profiles appear. However, other datasets would not be as well-behaved. Extending our influence measure to accommodate non-immediate influence is an important step towards implementing our results to other classification domains.

Finally, our experimental results, while encouraging, are illustrative rather than informative: they tell us that Google's display ads algorithm is clever enough to assign Spanish ads to Spanish speakers. Our experimental results enumerate the number of *displayed ads*; this is not necessarily indicative of

[2] `google.com/settings/ads`

users' clickthrough rates. Since our users are virtual entities, we are not able to measure their clickthrough rates; a broader experiment, where user profiles correspond to actual human subjects, would provide better insights into the effects user profiling has on display advertising.

Acknowledgments: This research was partially supported by the National Science Foundation (NSF) under grants CCF-0424422, CNS-1064688, CCF-1215883 and IIS-1350598, and by a Sloan Research Fellowship. The authors thank Joe Halpern for his insights regarding our results, as well as the inputs from IJCAI reviewers, which have substantially improved the paper.

References

[Bachrach *et al.*, 2010] Y. Bachrach, E. Markakis, E. Resnick, A.D. Procaccia, J.S. Rosenschein, and A. Saberi. Approximating power indices: theoretical and empirical analysis. *Autonomous Agents and Multi-Agent Systems*, 20(2):105–122, 2010.

[Balebako *et al.*, 2012] R. Balebako, P. Leon, R. Shay, B. Ur, Y. Wang, and L. Cranor. Measuring the effectiveness of privacy tools for limiting behavioral advertising. In *Web 2.0 Workshop on Security and Privacy*, 2012.

[Banzhaf, 1965] J.F. Banzhaf. Weighted voting doesn't work: a mathematical analysis. *Rutgers Law Review*, 19:317–343, 1965.

[Barocas and Nissenbaum, 2014] S. Barocas and H. Nissenbaum. Big data's end run around procedural privacy protections. *Communications of the ACM*, 57(11):31–33, October 2014.

[Blum and Langley, 1997] A.L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271, 1997.

[Calders and Verwer, 2010] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[Chalkiadakis *et al.*, 2010] G. Chalkiadakis, E. Elkind, E. Markakis, M. Polukarov, and N.R. Jennings. Cooperative games with overlapping coalitions. *Journal of Artificial Intelligence Research*, 39:179–216, 2010.

[Chalkiadakis *et al.*, 2011] G. Chalkiadakis, E. Elkind, and M. Wooldridge. *Computational Aspects of Cooperative Game Theory*. Morgan and Claypool, 2011.

[Chockler and Halpern, 2004] H. Chockler and J.Y. Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

[Cohen *et al.*, 2005] S. Cohen, E. Ruppin, and G. Dror. Feature selection based on the shapley value. In *IJCAI'05*, pages 665–670, 2005.

[Datta *et al.*, 2014] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. Technical Report arXiv:1408.6491v1, ArXiv, August 2014.

[Dwork *et al.*, 2012] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *ITCS'12*, pages 214–226, 2012.

[Guha *et al.*, 2010] S. Guha, B. Cheng, and P. Francis. Challenges in measuring online advertising systems. In *ACM SIGCOMM IMC'10*, pages 81–87, 2010.

[Halpern and Pearl, 2005] Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part I: Causes. *The British journal for the philosophy of science*, 56(4):843–887, 2005.

[Kamiran and Calders, 2009] F. Kamiran and T. Calders. Classifying without discriminating. In *IC4'09*, pages 1–6, 2009.

[Kamishima *et al.*, 2011] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *ICDMW'11*, pages 643–650, 2011.

[Lehrer, 1988] E. Lehrer. An axiomatization of the banzhaf value. *International Journal of Game Theory*, 17(2):89–99, 1988.

[Luong *et al.*, 2011] B.T. Luong, S. Ruggieri, and F. Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *SIGKDD'11*, pages 502–510, 2011.

[Marichal and Mossinghoff, 2006] J.L. Marichal and M. J. Mossinghoff. Slices, slabs, and sections of the unit hypercube. *arXiv preprint math/0607715*, 2006.

[Pedreschi *et al.*, 2008] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *SIGKDD'08*, pages 560–568, 2008.

[Podesta *et al.*, 2014] J. Podesta, P. Pritzker, E.J. Moniz, J. Holdern, and J. Zients. Big data: Seizing opportunities, preserving values. Technical report, Executive Office of the President - the White House, May 2014.

[Shapley, 1953] L.S. Shapley. A value for n-person games. In *Contributions to the Theory of Games, vol. 2*, Annals of Mathematics Studies, no. 28, pages 307–317. Princeton University Press, 1953.

[Tian and Pearl, 2000] J. Tian and J. Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

[Tschantz *et al.*, 2014] M.C. Tschantz, A. Datta, A. Datta, and J. M. Wing. A methodology for information flow experiments. *CoRR*, abs/1405.2376, 2014.

[Wills and Tatar, 2012] C.E. Wills and C. Tatar. Understanding what they do with what they know. In *WPES'12*, pages 13–18, 2012.

[Young, 1985] H.P. Young. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14(2):65–72, 1985.

[Zick *et al.*, 2014] Y. Zick, E. Markakis, and E. Elkind. Arbitration and stability in cooperative games with overlapping coalitions. *Journal of Artificial Intelligence Research*, 50:847–884, 2014.