Why Interpretability in Machine Learning? An Answer Using Distributed Detection and Data Fusion Theory

Kush R. Varshney [1] Prashant Khanduri [2] Pranay Sharma [2] Shan Zhang [2] Pramod K. Varshney [2]

Abstract

As artificial intelligence is increasingly affecting all parts of society and life, there is growing recognition that human interpretability of machine learning models is important. It is often argued that accuracy or other similar generalization performance metrics must be sacrificed in order to gain interpretability. Such arguments, however, fail to acknowledge that the overall decision-making system is composed of two entities: the learned model and a human who fuses together model outputs with his or her own information. As such, the relevant performance criteria should be for the entire system, not just for the machine learning component. In this work, we characterize the performance of such two-node tandem data fusion systems using the theory of distributed detection. In doing so, we work in the population setting and model interpretable learned models as multi-level quantizers. We prove that under our abstraction, the overall system of a human with an interpretable classifier outperforms one with a black box classifier.

1. Introduction

“When you create a Human+AI team, the hard part isn’t the ‘AI’. It isn’t even the ‘Human’. It’s the ‘+’” (Case, 2018).

Nirenburg (2017) dichotomizes artificial intelligence (AI) systems into cognitive prostheses, ones intended to replace humans, and cognitive orthotics, ones intended to enhance human performance on tasks. Also known as intelligence augmentation, orthotic systems are intended to collaborate with humans, and as such, must be proficient both at the task at hand and at communicating with humans. Computer systems can communicate at rates on the order of billions of bits per second, but humans can only do so on the order of hundreds of bits per second (Lawrence, 2018). A strength of humans, however, is intuition and reasoning (Case, 2018). Thus, to consider an AI system successful as an augmentation for decision support, it must bring forth relevant information for the decision making task, but must also communicate at an appropriate rate and in a way that allows a human recipient of the information to tap his or her strengths of intuition and reasoning.

Arguments in recent debates have claimed that it is only the accuracy of machine learning models that matters, not their interpretability. However, taking this view ignores the fact that the overall system in high-stakes settings is a machine learning model communicating with a human who makes the final decision, and thus it is the accuracy of the overall system that is of relevance. Interpretable machine learning models are an appropriate means for communication between AI and human (Dhurandhar et al., 2017); the contribution of this paper is to abstractly model the overall system and theoretically show the system performance advantage of interpretable machine learning models over black box machine learning models.

In this paper, we consider the population setting (the limit as the number of samples goes to infinity, allowing access to the probability distributions of the data) and appeal to the theory of distributed detection and data fusion (Varshney, 1997). We take this approach because it represents the simplest setting to understand the phenomenon without being too simple. Examples of working in the population setting abound in the machine learning literature (Gretton et al., 2006; Ravikumar et al., 2007; Scott et al., 2013; Shender & Lafferty, 2013; Menon & Williamson, 2018). We also restrict ourselves to binary classification for simplicity, but there is nothing fundamentally different if we consider multicategory classification. This work should be differentiated from recent contributions that discuss hybrids of interpretable and black box models (Wang, 2018), because here, we are concerned with the hybrid of *humans* and models.

The specific way we model the classification system is as a two-node sensor network in a tandem architecture. The first node is the machine learning model that makes a lo-

[1] IBM Research, Yorktown Heights, New York, USA [2] Syracuse University, Syracuse, New York, USA. Correspondence to: K. R. Varshney <krvarshn@us.ibm.com>.

2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden. Copyright by the author(s).

cal observation, puts it through the Bayes optimal decision rule (i.e. computes the likelihood ratio statistic),[1] and transmits a quantized version of this statistic to the second node, the human. The human has an independent local observation which it fuses with the information received from the model node to produce the final decision. The quantizer restricted to two quantization levels is used to model a black box model that can only transmit its classification. A quantizer with more than two quantization levels is used to model an interpretable classifier; one can imagine decision trees, rule sets, local post hoc explanations, and other human interpretable model forms (Malioutov et al., 2017) as partitions of the decision space similar to the effect of quantization of the likelihood ratio.

We prove that the Chernoff information between the two likelihood functions (class-conditional probabilities) participating in the final human decision is greater for systems with more quantization levels. Via the Chernoff theorem, this implies that the Bayes performance of the system with more quantization levels is better. That is: interpretable models perform better than black box models.

Note that we do not intend to imply that more levels yields greater interpretability, but only that three or more levels is an interpretable regime. In reality, a very large number of quantization levels stops being a good model for an interpretable machine learning model because humans have limits to how much information they can process. Therefore, let us assume that we are not in the regime with a large number of levels. In addition, we note that the proposed stylized abstraction of interpretability does not differentiate between simply quantizing the output score of a black box classifier with probabilistic outputs (which is still uninterpretable) and a truly human interpretable classifier; a more extensive formulation is needed to capture this distinction and incorporate additional aspects of interpretability such as the ability to examine feature-specific errors and vagaries caused by dataset shift. Other limitations of this work are discussed in Section 5.

2. Problem Setup

Consider the binary classification problem in the population setting with two nodes collaborating via a tandem network illustrated in Figure 1. Let the features observed by the two nodes, X_1 and X_2, be conditionally independent given the class label $Y \in \{0, 1\}$ and governed by the likelihoods $f_{X_1|Y}(x_1 \mid Y = y)$ and $f_{X_2|Y}(x_2 \mid Y = y)$. Sensor 1, a model for the machine learning model, transmits $U = \gamma(X_1)$ to sensor 2, a model for the human, where $\gamma(\cdot)$ is the composition of two functions, the likelihood ratio:

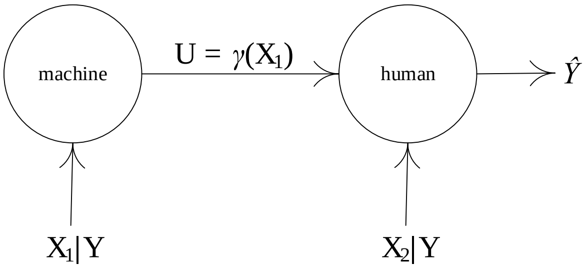

Figure 1. System model.

$$\Lambda(X_1) = \frac{f_{X_1|Y}(x_1 \mid y = 1)}{f_{X_1|Y}(x_1 \mid y = 0)},$$

and an optimal quantizer. Sensor 2 acts as a fusion center and bases its classification on both U and X_2. This classification rule $\hat{y}(U, X_2)$ is the globally Bayes optimal likelihood ratio test that thresholds

$$\Lambda(U, X_2) = \frac{f_{U,X_2|Y}(u, x_2 \mid y = 1)}{f_{U,X_2|Y}(u, x_2 \mid y = 0)}.$$

The quantizer has $k \geq 2$ levels. The case $k = 2$ models a black box and the case $k > 2$ models an interpretable model. Specifically,

$$U = \begin{cases} 1, & \Lambda(X_1) < b_1, \\ 2, & b_1 \leq \Lambda(X_1) < b_2, \\ \vdots & \vdots \\ k, & b_{k-1} \leq \Lambda(X_1) \end{cases} \tag{1}$$

where $\{b_1, b_2, \ldots, b_{k-1}\}$ are the quantization thresholds.

3. Performance Characterization

Our aim is to now show that the system having more quantization levels, i.e. larger k, has better classification performance. In service of that goal, we first provide a relevant inequality and prove that having more quantization levels leads to larger Chernoff information (or Chernoff divergence) (Chernoff, 1952) between the likelihood functions. Then we explicate how this relationship between Chernoff informations yields the conclusion of systems with interpretable classifiers performing better than systems with black box classifiers.

Lemma 1. *The following inequality is satisfied by posynomial functions f for $\lambda \in (0, 1)$:*

$$f\left(p_1^{1-\lambda} q_1^{\lambda}, \ldots, p_n^{1-\lambda} q_n^{\lambda}\right) \leq f\left(p_1, \ldots, p_n\right)^{1-\lambda} f\left(q_1, \ldots, q_n\right)^{\lambda} \tag{2}$$

[1] It is not obvious a priori that a local Bayes decision followed by quantization is optimal in this decision making architecture, but must be proved (Varshney, 1997; Zhu & Chen, 2013).

with equality if and only if $p_i = q_i$, $i = 1, \dots, n$.

Proof. This is Eq. 8 in Boyd et al. (2007). □

Remark. This geometric convexity inequality is a generalization of the arithmetic mean–geometric mean inequality.

Theorem 1. *Consider two learnable tandem networks as described above with different numbers of quantizer levels* k *and* k' *with* $k' > k$ *and corresponding quantized transmissions* U *and* U'*. Then, the following relationship among Chernoff informations holds:*

$$\begin{aligned} &C\left(f_{U',X_2|Y}(u',x_2 \mid y=1) \| f_{U',X_2|Y}(u',x_2 \mid y=0)\right) \\ &> C\left(f_{U,X_2|Y}(u,x_2 \mid y=1) \| f_{U,X_2|Y}(u,x_2 \mid y=0)\right). \end{aligned} \tag{3}$$

Proof. Since X_1 and X_2 are conditionally independent, for k-level quantization, we have:

$$\begin{aligned} &C\left(f_{U,X_2|Y}(u,x_2 \mid y=1) \| f_{U,X_2|Y}(u,x_2 \mid y=0)\right) \\ &\quad = C\left(f_{U|Y}(u \mid y=1) \| f_{U|Y}(u \mid y=0)\right) \\ &\qquad + C\left(f_{X_2|Y}(x_2 \mid y=1) \| f_{X_2|Y}(x_2 \mid y=0)\right). \end{aligned} \tag{4}$$

The second term in (4) does not depend on the quantization levels, so we focus only on the first term involving U. Recalling that U is a discrete random variable taking values $\{1, \dots, k\}$, this first term is given by

$$\begin{aligned} &C\left(f_{U|Y}(u \mid y=1) \| f_{U|Y}(u \mid y=0)\right) \\ &\qquad = -\log \min_{\lambda \in (0,1)} \sum_{j=1}^{k} p_j^{1-\lambda} q_j^{\lambda}, \end{aligned} \tag{5}$$

where $p_j = P(u = j \mid y = 1)$ and $q_j = P(u = j \mid y = 0)$. Similarly, for k'-level quantization, we have:

$$\begin{aligned} &C\left(f_{U'|Y}(u' \mid y=1) \| f_{U'|Y}(u' \mid y=0)\right) \\ &\qquad = -\log \min_{\lambda \in (0,1)} \sum_{i=1}^{k'} {p'_i}^{1-\lambda} {q'_i}^{\lambda}, \end{aligned} \tag{6}$$

where $p'_i = P(u' = i \mid y = 1)$ and $q'_i = P(u' = i \mid y = 0)$.

Without loss of generality, assume that the quantizer is a uniform quantizer. Then,

$$\begin{aligned} q_j &= P\left(\Lambda(X_1) \in \left[\tfrac{j-1}{k}, \tfrac{j}{k}\right) \mid y = 0\right), \\ p_j &= P\left(\Lambda(X_1) \in \left[\tfrac{j-1}{k}, \tfrac{j}{k}\right) \mid y = 1\right), \end{aligned}$$

for $j = 1, \dots, k$, and

$$\begin{aligned} q'_i &= P\left(\Lambda(X_1) \in \left[\tfrac{i-1}{k'}, \tfrac{i}{k'}\right) \mid y = 0\right), \\ p'_i &= P\left(\Lambda(X_1) \in \left[\tfrac{i-1}{k'}, \tfrac{i}{k'}\right) \mid y = 1\right), \end{aligned}$$

where $i = 1, \dots, k'$.

For k-level quantization, an interval $\left[\frac{r-1}{k}, \frac{r}{k}\right)$ contains $\rho = k'/k$ intervals of length $1/k'$.[2] Therefore, we have

$$\left[\tfrac{r-1}{k}, \tfrac{r}{k}\right) = \bigcup_{i=1}^{\rho} \left[\tfrac{\rho(r-1)+i-1}{k'}, \tfrac{\rho(r-1)+i}{k'}\right). \tag{7}$$

Then, using (a) the definition of q_j, (b) equation (7), (c) the disjoint property of intervals, and (d) the definition of q'_i:

$$\begin{aligned} q_r &\overset{(a)}{=} P\left(\Lambda(X_1) \in \left[\tfrac{r-1}{k}, \tfrac{r}{k}\right) \mid y = 0\right) \\ &\overset{(b)}{=} P\left(\Lambda(X_1) \in \bigcup_{i=1}^{\rho} \left[\tfrac{\rho(r-1)+i-1}{k'}, \tfrac{\rho(r-1)+i}{k'}\right) \mid y = 0\right) \\ &\overset{(c)}{=} \sum_{i=1}^{\rho} P\left(\Lambda(X_1) \in \left[\tfrac{\rho(r-1)+i-1}{k'}, \tfrac{\rho(r-1)+i}{k'}\right) \mid y = 0\right) \\ &\overset{(d)}{=} \sum_{i=1}^{\rho} q'_{\rho(r-1)+i}. \end{aligned} \tag{8}$$

Similarly,

$$p_r = \sum_{i=1}^{\rho} p'_{\rho(r-1)+i}. \tag{9}$$

Then, using (a) equation (5), (b) equations (8) and (9), (c) the geometric convexity inequality of Lemma 1, (d) collecting all the p'_i by summing r over $1, \dots, k$ and i over $1, \dots, \rho$, and (e) equation (6):

$$\begin{aligned} &C\left(f_{U|Y}(u \mid y=1) \| f_{U|Y}(u \mid y=0)\right) \\ &\overset{(a)}{=} -\log \min_{\lambda \in (0,1)} \sum_{r=1}^{k} p_r^{1-\lambda} q_r^{\lambda} \\ &\overset{(b)}{=} -\log \min_{\lambda \in (0,1)} \sum_{r=1}^{k} \left(\sum_{i=1}^{\rho} p'_{\rho(r-1)+i}\right)^{1-\lambda} \left(\sum_{i=1}^{\rho} q'_{\rho(r-1)+i}\right)^{\lambda} \\ &\overset{(c)}{<} -\log \min_{\lambda \in (0,1)} \sum_{r=1}^{k} \sum_{i=1}^{\rho} {p'_{\rho(r-1)+i}}^{1-\lambda} {q'_{\rho(r-1)+i}}^{\lambda} \\ &\overset{(d)}{=} -\log \min_{\lambda \in (0,1)} \sum_{i=1}^{k'} {p'_i}^{1-\lambda} {q'_i}^{\lambda} \\ &\overset{(e)}{=} C\left(f_{U'|Y}(u' \mid y=1) \| f_{U'|Y}(u' \mid y=0)\right). \end{aligned}$$

Step (c) is a strict inequality because the p'_i and the q'_i are different when the classification task is learnable. The result follows by reintroducing the second terms of equation (4). □

[2] For the sake of simplicity, we assume k' to be an integer multiple of k. The proof holds if that is not the case but requires additional bookkeeping.

Theorem 2. *The best achievable exponent in the Bayesian probability of error in a binary classification problem with class labels Y and features X is $C\left(f_{X|Y}(x \mid y=1) \| f_{X|Y}(x \mid y=0)\right)$.*

Proof. Known as the Chernoff Theorem, this is Theorem 11.9.1 in Cover & Thomas (2006). □

Theorem 3. *The probability of error in the tandem classification network described above with $k = 2$ quantizer levels is larger than the network with $k' > 2$ quantizer levels.*

Proof. This is a direct consequence of Theorem 1 and Theorem 2. □

Remark. This analysis makes no assumption about the relative quality of observations X_1 and X_2 made by the machine learning model and the human respectively. It continues to hold even if the two are very differently distributed (even on different variables), and the human features X_2 are very noisy—possibly relating to some intuition that is difficult to pin down and represent as data.

Remark. This analysis is for the Bayesian detection setting, which is the standard for supervised classification in machine learning. The detection theory literature is often oriented towards the Neyman–Pearson setting, which does occasionally also arise in machine learning (Rigollet & Tong, 2011). The current analysis can be repeated for the Neyman–Pearson paradigm with only minor changes: switching Chernoff information to Kullback–Leibler divergence, switching Lemma 1 to the log-sum inequality, and switching Theorem 2 to the Chernoff-Stein Lemma (Theorem 11.8.3 in Cover & Thomas (2006)).

4. Related Work in Distributed Detection and Estimation

The motivation for this study is to develop an understanding of the human–machine decision-making team in the presence of interpretable models, but it also provides a new contribution to the distributed detection and data fusion literature. Although two-node tandem sensor networks with quantization have been studied before, the analysis conducted in Section 3 has not been done.

Zhu & Chen (2013) investigated the problem of sufficiency-based data reduction in tandem fusion systems with quantization. They showed that quantizing the sufficient statistics achieves the same optimal inference performance as quantizing the raw observations. Their results applied to systems with conditionally independent observations and also to conditionally dependent observations under certain conditions. It is because of this result that we can equivalently quantize either $\Lambda(X_1)$ or X_1 in this work. This paper did not, however, characterize the difference in inference performance for different numbers of quantization levels as we do here.

Several works study the problem of whether the noisier sensor or the less noisy sensor in a two-sensor tandem fusion system should optimally perform the fusion and take the final decision (Akofor & Chen, 2013b; Zhu et al., 2013; Akofor & Chen, 2013a; Song et al., 2007; 2009). In many settings, it is the less noisy sensor that should optimally make the decision, but this is not universally true.[3]

Finally, Shen et al. (2012) and Shen et al. (2014) study a problem similar to ours with continuous Y, i.e. regression or estimation, and thus have characterizations hinging on Fisher information rather than Chernoff information as in the analysis herein. This work, like all of the others cited in this section, does not relate the analysis to interpretable machine learning and human decision making.

5. Limitations and Conclusion

In this paper, we have modeled the overall decision-making procedure involving humans and AI systems as one involving quantized communication from the AI to the human who makes the final decision. For analysis purposes, we have considered the population setting in which we can examine the probability distributions involved, thereby avoiding the complexities in analyzing the finite data sample regime. We have shown that interpretable AI (taken to be systems with more than two quantization levels) yields lower probability of error than black box AI (taken to be systems with two quantization levels).

One limitation of this work is that we have assumed that the two nodes, the human and the machine, have features that are independent conditioned on the label. However, it is quite reasonable to imagine that the two feature sets would be correlated, perhaps even strongly so and even statistically dependent. It is our conjecture that we can analyze the conditionally dependent case using Zhu & Chen (2013) as a starting point and following the approach that allows Shen et al. (2012) (conditionally independent measurements) to be extended to Shen et al. (2014) (conditionally dependent measurements), and that the main conclusion would not change.

Another limitation comes from working in the population setting. As discussed earlier, when we are in this setting, an optimal (scalar) quantizer of the sufficient statistic and an optimal (vector) quantizer of the raw features yield equivalent performance. The sufficient statistic represents a per-

[3] Scholars have raised this same question in discussing the collaboration of humans and AI in decision making. In this context, Kahneman recently stated (MIT IDE, 2018), “You can combine humans and machines, provided the machine has the last word!”

fect Bayes classifier and the optimal quantizer of the raw features also somehow captures perfect Bayes classification. Therefore, it is as if black boxes, post hoc interpretations, and directly interpretable models all have equivalent accuracies which is not necessarily true with finite training data. To extend the current analysis to the finite sample case, Nguyen et al. (2005) and Predd et al. (2006) may prove instructive; although they are for more general topologies than two-node tandems, which tend to introduce many simplifications and allow for analysis that may not otherwise be possible.

References

Akofor, Earnest and Chen, Biao. Interactive fusion in distributed detection: Architecture and performance analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4261–4265, Vancouver, Canada, May 2013a.

Akofor, Earnest and Chen, Biao. On optimal fusion architecture for a two-sensor tandem distributed detection system. In *IEEE Global Conference on Signal and Information Processing*, pp. 129–132, Austin, USA, December 2013b.

Boyd, Stephen, Kim, Seung-Jean, Vandenberghe, Lieven, and Hassibi, Arash. A tutorial on geometric programming. *Optimization and Engineering*, 8(1):67–127, March 2007.

Case, Nicky. How to become a centaur. *Journal of Design and Science*, February 2018.

Chernoff, Herman. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, December 1952.

Cover, Thomas M. and Thomas, Joy A. *Elements of Information Theory*. John Wiley & Sons, Hoboken, USA, 2006.

Dhurandhar, Amit, Iyengar, Vijay S., Luss, Ronny, and Shanmugam, Karthikeyan. A formal framework to characterize interpretability of procedures. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*, pp. 1–7, Sydney, Australia, August 2017.

Gretton, Arthur, Borgwardt, Karsten M., Rasch, Malte, Schölkopf, Bernhard, and Smola, Alex J. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, pp. 513–520, Vancouver, Canada, December 2006.

Lawrence, Neil. Natural and artificial intelligence. http://inverseprobability.com/2018/02/06/natural-and-artificial-intelligence, February 2018.

Malioutov, Dmitry M., Varshney, Kush R., Emad, Amin, and Dash, Sanjeeb. Learning interpretable classification rules with Boolean compressed sensing. In Cerquitelli, Tania, Quercia, Daniele, and Pasquale, Frank (eds.), *Transparent Data Mining for Big and Small Data*, pp. 95–121. Springer, Cham, Switzerland, 2017.

Menon, Aditya Krishna and Williamson, Robert C. The cost of fairness in binary classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, pp. 107–118, New York, USA, February 2018.

MIT IDE. Where humans meet machines: Intuition, expertise and learning. https://medium.com/mit-initiative-on-the-digital-economy/where-humans-meet-machines-intuition-expertise-and-learning-be639f00bade, May 2018.

Nguyen, XuanLong, Wainwright, Martin J., and Jordan, Michael I. Nonparametric decentralized detection using kernel methods. *IEEE Transactions on Signal Processing*, 53(11):4053–4066, November 2005.

Nirenburg, Sergei. Cognitive systems: Toward human-level functionality. *AI Magazine*, 38(4):5–12, Winter 2017.

Predd, Joel B., Kulkarni, Sanjeev R., and Poor, H. Vincent. Consistency in models for distributed learning under communication constraints. *IEEE Transactions on Information Theory*, 52(1):52–63, January 2006.

Ravikumar, Pradeep K., Liu, Han, Lafferty, John, and Wasserman, Larry. Spam: Sparse additive models. In *Advances in Neural Information Processing Systems 20*, pp. 1201–1208, Vancouver, Canada, December 2007.

Rigollet, Philippe and Tong, Xin. Neyman-Pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 12:2831–2855, October 2011.

Scott, Clayton, Blanchard, Gilles, and Handy, Gregory. Classification with asymmetric label noise: Consistency and maximal denoising. In *Proceedings of the Conference on Learning Theory*, pp. 489–511, Princeton, USA, June 2013.

Shen, Xiaojing, Varshney, Pramod K., and Zhu, Yunmin. Robust distributed maximum likelihood estimation with quantized data. arXiv:1208.4161v1, August 2012.

Shen, Xiaojing, Varshney, Pramod K., and Zhu, Yunmin. Robust distributed maximum likelihood estimation with dependent quantized data. *Automatica*, 50(1):169–174, January 2014.

Shender, Dinah and Lafferty, John. Computation-risk tradeoffs for covariance-threshold regression. In *Proceedings of the International Conference on Machine Learning*, pp. 756–764, Atlanta, USA, June 2013.

Song, Enbin, Zhu, Yunmin, and Zhou, Jie. Some progress in sensor network decision fusion. *Journal of Systems Science and Complexity*, 20(2):293–303, June 2007.

Song, Enbin, Shen, Xiaojing, Zhou, Jie, Zhu, Yunmin, and You, Zhisheng. Performance analysis of communication direction for two-sensor tandem binary decision system. *IEEE Transactions on Information Theory*, 55(10):4777–4785, October 2009.

Varshney, Pramod K. *Distributed Detection and Data Fusion*. Springer-Verlag, Secaucus, USA, 1997.

Wang, Tong. Hybrid decision making: When interpretable models collaborate with black-box models. arXiv:1802.04346, February 2018.

Zhu, Shengyu and Chen, Biao. Data reduction in tandem fusion systems. In *Proceedings of the IEEE China Summit & International Conference on Signal and Information Processing*, pp. 602–606, Beijing, China, July 2013.

Zhu, Shengyu, Akofor, Earnest, and Chen, Biao. Interactive distributed detection with conditionally independent observations. In *Proceedings of the IEEE Wireless Communications and Networking Conference*, pp. 2531–2535, Shanghai, China, April 2013.