better accuracy than the scores used by both $\|\Phi(x)\|$ and iNNE. This is because the characterisation provides an effective reference for point anomaly detection, robust to contamination of anomalies in a dataset. This robustness is important when using points in a dataset which contains anomalies to build a model (that consists of hyperspheres used in $\widehat{\mathcal{K}}_I$, $\|\Phi(x)\|$ and iNNE.)

Since $\widehat{\mathcal{K}}_I$, iNNE [2] and iForest [8] have same time complexity $O(nt\psi)$, they have approximately the same runtime on the largest dataset http, as shown in Table 3.

In a nutshell, existing isolation-based anomaly detectors, i.e., iNNE and iForest, employ a score similar to the norm $\|\Phi(x)\|$. This is an interesting revelation because isolation-based anomaly detectors were never considered to be related to a kernel-based method before the current work. The power of isolation-based anomaly detectors can now be directly attributed to the norm of the feature map $\|\Phi(x)\|$ of Isolation Kernel.

From another perspective, $\widehat{\mathcal{K}}_I$ can be regarded as a member of the family of Isolation-based anomaly detectors [2, 8]; and iForest [8] has been regarded as one of the state-of-the-art point anomaly detectors [1, 6]. iNNE [2] is recently proposed to be an improvement of iForest. **$\widehat{\mathcal{K}}_I$ is the only isolation-based anomaly detector that is also a kernel based anomaly detector.**

The improvement of iNNE over iForest is mainly due to the use of a better isolating mechanism—overcoming four weaknesses of iForest [2]. The improvement of $\widehat{\mathcal{K}}_I$ over iNNE is mainly due to the distributional characterisation—it also contributes to $\widehat{\mathcal{K}}_I$ being the most robust isolation-based anomaly detector against contamination of anomalies in the training set.

In addition, the proposed method is the only detector that makes use of distributional kernel for point anomaly detection. While existing detector OCSMM [11] employs a distributional kernel, it is a group anomaly detector, not a point anomaly detector.

8 CONCLUSIONS

We show that Isolation Distributional Kernel addresses two key issues of kernel mean embedding, where the kernel employed has: (i) a feature map with intractable dimensionality which leads to high computational cost; and (ii) data independency which leads to poor detection accuracy in anomaly detection.

Our theoretical analyses reveal that a point kernel must be both data dependent and characteristic in order to produce an effective anomaly detector. Gaussian kernel, being characteristic but data independent, does not have the sufficient conditions.

We introduce a new Isolation Kernel and establish the geometrical interpretation of its feature map. This interpretation provides the insight that point anomalies and normal points are mapped into distinct regions in the feature space. This is the source of the power of $\widehat{\mathcal{K}}_I$. We also reveal that the distributional characterisation makes $\widehat{\mathcal{K}}_I$ robust to contamination of anomalies in a dataset.

Our evaluation shows that $\widehat{\mathcal{K}}_I$, without explicit learning, produces better detection accuracy than existing key kernel-based methods in detecting point anomalies, while achieving short testing and training times. In contrast, $\widehat{\mathcal{K}}_G$ and OCSVM, which employ Gaussian kernel, have lower detection accuracy in most datasets, and run up to three orders of magnitude slower in large datasets.

9 ACKNOWLEDGEMENTS

This research was supported by the National Key R&D Program of China (2018YFB1004300), NSFC (61921006), 111 Program (B14020) and JST CREST (JPMJCR1666).

REFERENCES

[1] Charu C. Aggarwal and Saket Sathe. 2017. *Outlier Ensembles: An Introduction.* Springer International Publishing.
[2] Tharindu R. Bandaragoda, Kai Ming Ting, David Albrecht, Fei Tony Liu, Ye Zhu, and Jonathan R. Wells. 2018. Isolation-based Anomaly Detection using nearest neighbour ensembles. *Computational Intelligence* 34, 4 (2018), 968–998.
[3] Vasili Baranau and Ulrich Tallarek. 2014. Random-close packing limits for monodisperse and polydisperse hard spheres. *Soft Matter* 10 (2014), 3826–3841.
[4] Liefeng Bo and Cristian Sminchisescu. 2009. Efficient Match Kernels Between Sets of Features for Visual Recognition. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*. 135–143.
[5] Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research* (2006), 1–30.
[6] Andrew Emmott, Shubhomoy Das, Thomas G. Dietterich, Alan Fern, and Weng-Keen Wong. 2016. A Meta-Analysis of the Anomaly Detection Problem. *CoRR* abs/1503.01158 (2016).
[7] Fukunaga Keinosuke. 1990. *Introduction to Statistical Pattern Recognition.* Academic Press, Chapter 6, 268–270.
[8] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *Proceedings of the IEEE International Conference on Data Mining*. 413–422.
[9] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Scholkopf. 2012. Learning from Distributions via Support Measure Machines. In *Advances in Neural Information Processing Systems*. 10–18.
[10] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard SchÃűlkopf. 2017. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends in Machine Learning* 10 (1âĂŞ2) (2017), 1–141.
[11] Krikamol Muandet and Bernhard Schölkopf. 2013. One-class Support Measure Machines for Group Anomaly Detection. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*. 449–458.
[12] Cameron Musco and Christopher Musco. 2017. Recursive sampling for the nystrom method. In *Advances in Neural Information Processing Systems*. 3833–3845.
[13] Xiaoyu Qin, Kai Ming Ting, Ye Zhu, and Vincent Cheng Siong Lee. 2019. Nearest-Neighbour-Induced Isolation Similarity and Its Impact on Density-Based Clustering. In *Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence*.
[14] Ali Rahimi and Benjamin Recht. 2007. Random Features for Large-scale Kernel Machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*. 1177–1184.
[15] Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. 2001. Estimating the Support of a High-Dimensional Distribution. *Neural Computing* 13, 7 (2001), 1443–1471.
[16] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. 2007. A Hilbert Space Embedding for Distributions. In *Algorithmic Learning Theory*, Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto (Eds.). Springer, 13–31.
[17] Chaoming Song, Ping Wang, and Hernan A. Makse. 2008. A phase diagram for jammed matter. *Nature* 453, 7195 (2008), 629–632.
[18] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. 2010. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research* 11 (2010), 1517–1561.
[19] Dougal J. Sutherland. 2016. *Scalable, Flexible and Active Learning on Distributions.* PhD Thesis, School of Computer Science, Carnegie Mellon University.
[20] Kai Ming Ting, Yue Zhu, and Zhi-Hua Zhou. 2018. Isolation Kernel and its effect on SVM. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2329–2337.
[21] Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10, 2 (2009), 207–244.
[22] Christopher K. I. Williams and Matthias Seeger. 2001. Using the Nyström Method to Speed Up Kernel Machines. In *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp (Eds.). 682–688.
[23] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. 2002. Distance Metric Learning, with Application to Clustering with Side-information. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*. 521–528.
[24] Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. 2012. Nyström Method vs Random Fourier Features: A Theoretical and Empirical Comparison. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12)*. Curran Associates Inc., USA, 476–484.
[25] Pourya Zadeh, Reshad Hosseini, and Suvrit Sra. 2016. Geometric mean metric learning. In *International Conference on Machine Learning*. 2464–2471.