Résumé sur la détection d'anomalies

DING Wensi

June 15, 2016

1 Modèle Gaussien

Algorithme:

1. choisir des variables qui peuvent servir à distinguer les points normaux et les points anormaux. Données: $\{x^1, x^2, \cdots, x^m\}$ avec $x^i \in R^n$, chaque x^i_j représente une variable choisie

2. construire une distribution gaussien séparément pour chaque x_j, c'est-à-dire, calculer la moyenne μ_j et la variance σ_j^2 à partir des x^i_j

3. pour un nouvelle donnée x, calculer $p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2)$, et si $p(x) > \epsilon$ alors cette donnée est prédit comme normale, sinon elle est anormale

Distribution de données:

training set: 60% de données négatives

cross validation set: 20% de données négatives et 50% de données positives

test set: 20% de données négatives et 50% de données positives

Pour que ce modèle fonctionne bien, nous devons avoir des variables et aussi un seuil ϵ bien choisies. Pour cela, il est indispensable d'avoir un système d'évaluation: F1-score.

Il est utile de vérifier les variables choisies sont bien gaussiennes par dessiner les histogrammes.

Il est possible que certaines variables ne sont pas indépendantes. Dans ce cas, cet algorithme ne fonctionne pas très bien. Nous pouvons améliorer cela par créer des nouvelles variables en utilisant les variables indépendantes ou par utiliser une multi-variables gaussien distribution.

Insupervised learning: petit nombre de données positives et capable de détecter des anomalies qui n'existent pas avant

Supervised learning: grand nombre de données positives et négatives et incapable de détecter des anomalies qui n'existent pas avant

Modèle Gaussien Simple	Modèle Gaussien Multi-variables
détecter les variables indépendantes manuellement	automatique
demander moins de ressource pour computation	cher
marcher bien quand le nombre de données est petit	non

2 Les grandes catégories

2.1 Quelques détails

Les différents **types d'anomalie**: des anomalies simples (point anomalies), des anomalies complexes (anomalies contextuelles, anomalies collectives)

Contraintes principales: type de données fournies (continue ou discret), label pour chaque donnée est fourni ou pas, et des contraintes spéciales pour certains domaines associés

Selon la disponibilité des labels des données,la détection peut fonctionner dans le mode d'**apprentissage supervisé**, **apprentissage semi-supervisé** ou **apprentissage non-supervisé**.

L'**output** de l'algorithme de détection d'anomalies est soit un score, soit un label.

Domaine d'application: intrusion detection, fraud detection, Medical and Public Health Anomaly Detection, industrial Damage Detection, image Processing, anomaly Detection in Text Data, sensor Networks etc.

2.2 Classification based

Apprentissage Supervisé

Hypothèse: l'algo peut apprendre à distinguer les différents class à partir des données fournies

Deux étapes: training phase et testing phase

L'idée principale: construire un modèle qui apprend à caractériser différents class normaux (un ou plusieurs) et si la nouvelle donnée n'appartient à aucun class normaux, alors elle est un point anormal.

Algo principaux: Neural Networks, Bayesian Networks, Support Vector Machines, Rule Based

Complexité dépend de l'algo utilisé pour la classification

Avantage: testing process est rapide

Désavantage: l'information sur les labels n'est souvent pas fournie; L'output est un label mais pas un score, mais ceci peut être résolu facilement.

2.3 Plus Proche Voisin based

Apprentissage non-supervisé ou semi-supervisé

Hypothèse: Les points normaux restent proches, au contraire, les points anormaux sont loin de leur plus proche voisin. Donc cette approche ne marchera pas quand le nombre de points anormaux aui sont proches est grand.

Une distance entre deux données doit être définie.

L'idée principale:

1. Le score est fixé par la distance de son k ième plus proche voisin, un seuil est imposé pour faire la décision.
2. Quand la densité de différents ensembles varie, la critère doit être la densité relative (LOF: le rapport entre la moyenne de la densité locale de ses voisins et sa propre densité locale).

Algo principaux varient dans les trois aspects:

1. définition de la distance pour les différents types de variables
2. définition du score
3. amélioration sur la complexité

Complexité: $O(N^2)$. Certains algorithmes ont une meilleure complexité à condition de petite dimension ou de s'intéresser seulement aux top anomalies

Avantage: Apprentissage non-supervisé ou semi-supervisé

Désavantage: Hypothèse n'est pas nécessairement vraie; Complexité pour le testing process reste haute; Distance peut être difficile à définir; Pour les données de grande dimension, cette méthode ne marche pas

2.4 Clustering based

Apprentissage non-supervisé ou semi-supervisé

Trois Hypothèse:

- point normal appartient à un des clusters, point anormal n'appartient à aucun cluster
- point normal reste proche de la centroid du cluster plus proche, point anormal est loin de la centroid du cluster plus proche
- point normaux forment un cluster dense, point anormaux forment un petit cluster

Les deux premières hypothèses ne marchent pas pour le cas où les points anormaux forment aussi un petit cluster. Les trois hypothèses entraînent trois différents définition pour le score.

Complexité: Dépend de l'algorithme de clustering (de linéaire à quadratique)

Avantage: Apprentissage non-supervisé; testing process est rapide

Désavantage: Hypothèse n'est pas nécessairement vraie; Performance dépend de l'algorithme de clustering; Pour les données de grande dimension, cette méthode ne marche pas

2.5 Méthode Statistique

Apprentissage non-supervisé ou semi-supervisé

Hypothèse: Pour un modèle stochastique, les points normaux tombent dans un région de grande probabilité et les points anormaux tombent dans un région de petite probabilité

Selon les connaissances que nous avons, deux approches est possibles: paramétrique ou non-paramétrique.L'approche paramétrique: gaussian based, regression based, mixture of distribution based; L'approche non-paramétrique: histogramme based, kernel function based.

Complexité: Dépend du modèle statistique que nous utilisons (de linéaire à quadratique)

Avantage: Si l'estimation de distribution est robuste aux données anormales, cette méthode peut marche avec le mode d'apprentissage non-supervisé; testing process est rapide

Désavantage: Un modèle convenable n'existe pas nécessairement pour les données de grande dimension; Difficile à choisir un modèle statistique convenable

2.6 Information Theory based

Apprentissage non-supervisé

Hypothèse: Les points anormaux augmentent l'irrégularité de la contenu d'information du data set

L'idée principale: trouver un sous-ensemble qui entraîne la plus grande différence de la complexité d'information pour le data set total et le data set moins ce sous-ensemble

Complexité: Exponentiel, quelques amélioration possible pour un temps linéaire

Avantage: Apprentissage non-supervisé

Désavantage: Souvent cette méthode ne marche que pour un grand nombre de points anormaux; Difficile pour obtenir un score; Testing process est lent

2.7 Spectral Theory based

Apprentissage non-supervisé ou semi-supervisé

Hypothèse: Les points normaux et anormaux deviennent distinguable pour un sous-espace avec moins de dimension

La mission principale est de trouver un sous-espace qui fonctionne.

Complexité: PCA (linéaire par rapport à la taille de données, quadratique pour la dimension)

Avantage: Apprentissage non-supervisé; Fonctionner pour les données de grande dimension

Désavantage: Un tel sous-espace n'existe pas nécessairement; Testing process est lent

2.8 Anomalies Contextuelles

La décision dépend de la contexte. Donc les variables sont séparées en deux parties: variable contextuelle et variable indicatives.

Deux approches: transformer le problème à un problème d'anomalie simple (facile quand les variables contextuelles sont évidentes); construire un modèle à partir des structures de données (prédire le comportement pour chaque donnée, et le comparer avec le comportement réel)

Complexité: Pour la première approche, ça dépend de l'algorithme utilisé pour la détection d'anomalie simple; Pour la deuxième approche, souvent elle coûte plus chère, mais elle est plus rapide pour le testing process

Avantage: Parfois c'est une approche plus naturelle

Désavantage: Difficile à trouver des variables contextuelles convenables

2.9 Anomalies Collectives

L'occurrence de certaine combinaison est considérée comme anormale. Chaque point de cette combinaison n'est pas nécessairement anormal.

Deux catégories: anomalie séquentielle et anomalie spatiale

Anomalie séquentielle:

- trouver une séquence anormale parmi un ensemble de séquences
- trouver une sous-séquence anormale pour une séquence longe
- déterminer la fréquence d'un certain pattern est normal ou pas pour une séquence

Anomalie spatiale: Principalement pour le domaine de traitement d'images