SCIENCE DES DONNEES

Daniel MARANDICI[1*]

[1]*Université Technique de Moldavie, Faculté Ordinateur, Informatique et Microélectronique, Département Génie Logiciel et Automatique, gr.FI-201, Chișinău, Moldova*

*Autorul corespondent: Marandici Daniel, daniel.marandici@isa.utm.md

Résumé. *Au cours des dernières décennies les données ont explosé et l'augmentation rapide de leur quantité dans le cyberespace a amené l'humanité dans l'ère des mégadonnées. La signification et la valeur des données a évolué considérablement en transformant une science dans une précieuse industrie.*

Mots-clés : *statistiques, informatique, analyse.*

Introduction. En 1996, le terme de science des données a été introduit pour la première fois dans le titre d'une conférence sur les statistiques (IFCS, "Data Science, classification, and related methods") [1]. Au début la science des données a eu une forme pure mathématique, en se basant seulement sur tests des statistiques, création d'hypothèses et compréhension de données [2]. Après que l'ordinateur a été inventé nous avons commencé a utiliser et traiter des données. En tant que les ordinateurs ont devenu de plus en plus puissants et performants, la science de données a évolué dans une industrie réelle. Ces changements ont résulté dans des nouveaux défis pour les technologies de ce secteur. À ce moment, la formule du Cao [3] définit parfaitement les domaines concernés en science des données :

$$\textit{Science des données} = (\textit{statistiques} + \textit{informatique} + \textit{calcul} + \textit{communication} + \textit{sociologie} + \textit{gestion}) / (\textit{données} + \textit{environnement} + \textit{réflexion}) \quad (1)$$

En cette formule, la sociologie représente la psychologie humaine et / (données + environnement + réflexion) veut dire que toutes les sciences mentionnées se fondent sur la base de données. On peut constater que le but d'utilisation des données s'est modifié aussi. Au départ seulement pour générer des faits possibles jusqu'à présent, quand la science des données ensemble avec l'intelligence artificielle étudient les évènements politiques, économiques, analysent chaque mouvement d'utilisateur partout sur internet pour s'améliorer constamment et donner des meilleures prémisses.

Les processus de base de la science des données

1. Pour collecter les données les programmeurs expérimentés utilisent la conception d'expériences (CDE), qui est essentielle pour obtenir des informations qualitatives quand il y a des facteurs qui peuvent perturber les données. Dans le cas où ces facteurs sont identifiés et leurs conséquences sont minimisés il y a encore le milieu environnemental qui est impossible de contrôler. Mais la CDE a prouvé son efficacité et elle est le moyen primaire pour fournir de nouvelles données, pour optimiser les algorithmes et perfectionner les méthodes d'analyse eux-mêmes. La solution pratique pour équilibrer les lacunes de données causées par erreurs est la simulation [4]. La CDE et la simulation sont les piliers fondamentaux de cette science.
2. Exploration de données est un autre procès essentiel. En anglais « data mining » est indispensable pour mettre en pratique les méthodes analytiques de la science de données. La distribution, une notion importante pour représenter la fluctuation des données.

3. Analyse statistique des données est le lien entre les processus précédents et suivants. Quelques exemples des méthodes statistiques :
    I. Vérification des hypothèses est utile pour tester les questions possibles. Les théories pour le futur sont essayées sur les données existantes, de cette façon les compagnies planifient leurs activités mieux en sachant les cas réels du futur en contexte actuel. D'autre part ce méthode permet d'observer les fautes et les résultats de certaines décisions prises en avant.
    II. Pour trouver et prévoir les sous-populations à partir des données les spécialistes appliquent les méthodes de classification, aussi appelées regroupement (clustering , EN) [5] aux facteurs influents. Mais le succès d'attendre aux lois exige la mise en œuvre d'algorithmes parfaitement optimisés avec la complexité temporelle bas, ce qui n'est pas facile de tout à fait pour développeur si on a peu d'expérience. En général on travaille avec mégadonnées et si l'algorithme est faible ou pas optimisé au maximum, le temp de calcul de l'analyse statistique complexe est énorme et en conséquence pas pratique, réel a implémenter.
    III. Les méthodes de régression, en particulier la régression linéaire est l'outil principal pour identifier des relations entre caractéristiques globaux et locaux sur une valeur spécifique. Simplifié, la régression trouve l'impact d'un événement large sur un autre en minorité et vice-versa pour un élément concret.
4. La modélisation statistique commence à définir une conception à l'ensemble des données. Une manière de faire c'est d'utiliser les graphs [6] et les chaînes pour connecter l'interaction des deux facteurs. Un autre moyen est le suivant : si on a obtenu un graphique, un tracé pour les données, on s'approche de lui comme d'une fonction mathématique. A cet instant, on peut appliquer les équations différentielles à un modèle approximatif pour recevoir des informations qualitatives. En fin la hiérarchie est une autre solution pour structurer les modèles globaux et locaux et analyser en temps les bouleversements.
5. En avant de présenter les résultats, valider et sélectionner les modèles est critique pour obtenir ceux avec la meilleure puissance prédictive et performance. En ce cas la comparaison, les expériences de perturbation, la méta-analyse et la sélection des modèles sont les techniques pratiquées par développeurs en derniers années.
6. Le dernier processus et le plus important de la science des données est la représentation et les reports. En cette étape est nécessaire une visualisation des structures obtenues et une communication simplifiée des résultats avant de déploiement. Dans le report on décrit le but, les techniques pratiquées, les méthodes implémentées, les analyses réalisées, la marge d'erreur possible, on interprète les résultats [7] en termes adaptés pour les destinataires et a la fin les conclusions et les recommandations. C'est le dernier moment et le plus significatif quand la science des données est impliquée auparavant que les entreprises finissent leur plan et passent aux actions.

Les défis

Le problème primordial est la confiance en donnés reçues. Comment on peut savoir si les informations sont véridiques ? Comment on peut éliminer celles fausses ? Si elles sont mélangées comment on les gère ? Souvent la compétition entre les entreprises amène aux opinions et commentaires mauvaises et pas réalistes. Par conséquence les nouveaux clients, utilisateurs sont affectés par l'image négative d'entreprise et la crédibilité des données collectées est baissée. En observant le rythme de la croissance des réseaux sociaux, ces challenges deviennent de plus en plus sévères.

Un autre défi est l'explosion des données [8] partout les domaines. En présent la science attente des nombreuses découvertes et le domaine IT interfère et devient une partie composante de tous les secteurs. Par conséquence on a besoin constamment des nouvelles approches des données, parce que les méthodes traditionnelles ne sont pas pertinentes. À ce moment on prenne conscience de la valeur de la science de données et ça exige d'explorer en permanence et identifier des stratégies actuelles.

Actuellement les questions qui nécessitent encore d'études et analyses sont les suivantes :

- Comment on peut trouver des données utiles dans le cyberespace ?
- Comment on peut obtenir des connaissances tirées des données ?

Ça requière d'observer les données d'un point de vue nouveaux, différent.

Importance et impact

En termes simples le but de la science des données est de faire prédictions et analyser les processus dans le monde qui nous entoure. Si on parle des compagnies d'informatique, la majorité qui ont faites l'analyse des données et puis on actionnées en se basant sur les résultats obtenus ont eu une croissance formidable du succès des projets et portfolios. Analyser les données est indispensable pour la prospérité d'entreprise à long terme. Les statistiques des projets aident les managers à observer les fautes pour les diminuer en futur. Pour la nature et environnement la science des données est appliquée en vue de révéler des nouvelles règles de comportement d'animaux et d'observer minutieusement l'impact des activités humaines sur le milieu naturel. En plus, les informations collectées sur les réseaux sociaux sont utiles pour mieux comprendre la psychologie humaine. On peut affirmer avec confiance que partout où il y a des produits technologiques là il y a des données. Par conséquence, le domaine d'application de cette science n'a pas des limites, parce qu'en présent l'industrie IT est omniprésente.

Les sciences et technologies concernées

- Langages de programmation
    1. Python
    2. R
    3. Julia
- Acquisition et exploration des données
    1. Talend
    2. Mozenda
    3. Octoparse
- Stockage et gestion des données
    1. MySQL
    2. SQL
- Sécurité cybernétique
- Analyse et visualisation des données
    1. Plotly
    2. Tableau
    3. Qlik
    4. AnyChart
    5. Google Charts
    6. Webix
- Apprentissage automatique
- Intelligence artificielle
    1. Skicit-learn
    2. SciPi
    3. TensorFlow
    4. Keras

- Mathématiques
    1. Data Melt
    2. MATLAB
- Statistiques
- Bio-informatique

Conclusion

Si on compare la science des données avec la science informatique, elle est sous-estimée, mais on ne doit pas mettre une supérieure à l'autre. Seulement ensemble avec l'informatique, mathématiques et statistiques on va identifier avec succès des solutions, parce qu'aujourd'hui toutes les sciences sont liées les unes aux autres. En plus, ce domaine est toujours en changement, en évolution et ça lui permet d'être actualisé, d'avoir une demande ici et maintenant, caractéristique pas spécifique à des nombreuses sciences.

Comment on a mentionné avant, l'application de la science des données et sans limites, par exemple en santé, en sports, pour gouvernement et pour transport public. Tous les efforts concentrés en ces secteurs ont le but d'améliorer et faciliter la vie d'humanité, pas seulement d'aider les entreprises comment on a pu avoir la première impression. Un bonne exemple et astucieux concernant l'environnement est que l'intelligence artificielle à optimiser les itinéraires de transport pour diminuer la pollution et économiser les gas-oil basée sur les informations collectées.

Finalement, on peut conclure que la science des données est différente des sciences actuelles et sera une orientation significative et prometteuse à l'avenir. On croit qu'en futur la science des données deviendra un nouveau type de science, comment on a à ce moment les sciences culturelles et sociaux.

Bibliographie :

1. PRESS, G., *A Very Short History of Data Science* [online], 2013. [accesat 31.01.2021]. Disponibil: https://www. forbescom/sites/gilpress/2013/05/28/a-very-short-history-ofdata-science/#5c515ed055c
2. ZHU, Y., XIONG, Y., *Towards Data Science* [online], 2015. Pp 1-7 [accesat 31.01.2021]. Disponibil : http://dx.doi.org/10.5334/dsj-2015-008CO
3. CAO, L., *Data science: a comprehensive overview*. ACM Comput. Surv. (2017). https://doi.org/10.1145/3076253
4. BERGER, R.E., *A scientific approach to writing for engineers and scientists*. IEEE PCS Professional Engineering Communication Series IEEE Press, 2014
5. HENNIG, C., MEILA, M., MURTAGH, F., ROCCI, R. : *Hand book of Cluster Analysis. Chapman & Hall*, 2015.
6. KOLLER, D., FRIEDMAN, N.: *Probabilistic Graphical Models: Principles and Techniques.* MIT press, Cambridge (2009).
7. WEIHS, C., ICKSTADT, K., *Data Science: the impact of statistics* [online], 2018, 02. Pp 189-194. [accesat 30.01.2021]. Disponibil: https://doi.org/10.1007/s41060-018-0102-5
8. ZHU, Y. Y., ZHONG, N., & XIONG, Y. Data Explosion, Data Nature and Dataology. In *Proceedings of International Conference on Brain Informatics,* 2009.