



UNIVERSITÉ
JEAN MONNET
SPÉCIALITÉ WEB
INTELLIGENCE

TSINGHUQ
UNIVERSITY
DEPARTMENT OF ELECTRONIC
ENGINEERING

STAGE EN ENTREPRISE: STAGE DE FIN D'ÉTUDE

Rapport de Stage Année 2013-2014

Fouille de Donnée dans la domaine de
télécommunication

Auteur :
Wenyi WANG

Tuteur de stage en entreprise:
Vice directeur de labo NGN:
yongfeng HUANG
Tuteur de l'université:
Amaury HABRARD

De 20 Février 2014 à 20 Juillet 2014

Table des matières

Remerciements	3
1 Résumé	1
2 Introduction	1
2.1 Introduction du CMCC	2
2.2 La crise du CMCC	2
2.3 L'optimisation du réseau	5
2.4 Introduction du laboratoire	6
2.5 Objectif du projet	7
3 Introduction de l'industrie de la télécommunication	8
3.1 L'evolution des normes de téléphonie mobile	8
3.1.1 La premier génération	9
3.1.2 La deuxième génération	9
3.1.3 La troisième génération	10
3.1.4 La quatrième génération	10
3.2 Le réseau LTE	11
3.2.1 La structure du réseau LTE	12
4 Les solution existant	14
5 Le présentation de notre solution	17
5.1 Clustering	19
5.1.1 La distance	19
5.1.2 La validité de clustering	20
5.1.3 La silhouette Cœfficient	21
5.2 Règles d'association	23
5.3 La mise en œuvre	24
5.3.1 Le logiciel utilisés	24

5.3.2	Introduction des données	25
5.3.3	Prétraitement de données	26
5.3.4	Les caractéristique du donnée	29
6	La mise en œuvre	29
6.1	La valeur optimal de k	29
6.1.1	La somme d'erreur carrés	30
6.1.2	La silhouette Coefficient	30
	Conclusion	35
	Références	36

Remerciements

Tout d'abord, je tiens à remercier Amaury Habrard et tous les enseignants de la Spécialité Web Intelligence de l'Université Jean Monnet, aussi les enseignants de Télécom Saint-Etienne et L'école nationale supérieure de Saint-Etienne, qui m'a aidé lors de ces deux années de étude.

Je remercie également M.Yongfeng HUANG pour avoir accepter diriger cette stage, il m'a beaucoup conseillé, et les discussions que l'on a pu avoir se sont toujours révélées très intéressantes et instructives.

Je souhaite également adresser mes remerciement à Zheng YANG, Lindong WEI et xian WU ainsi que tout les membres du laboratoire de Next generation Network(NGN) pour m'avoir soutenu, encouragé et conseillé tout au long de ce stage.

Je tiens à montrer tout ma gratitude envers toutes les personnes qui ont pu m'aider, m'encourager, me soutenir, me remotiver pendant ces années de travail.

1 Résumé

Pendant ces quatre mois de stage, notre groupe de recherche travaille avec les employés de CMCC (China Mobile Communications Corporation) . Le objectif du sujet est: utilise les technique de Fouille de données, étude les données fournir par le CMCC, et trouve les relation entre les données et les défaut du système 4G. Nous avons fait plusieurs tentatives pour trouver les résultats, et on a utilise différents logiciel, j'ai utilisé le R, et mon collègue utilise Mathlab, nous avons utilisé plusieurs algorithmes (Clusterring, PCA, Association rules, Ajustement). Mais à la fin, nous avons trouvé que à cause des défaut dans la système d'acquisition, les données ne sont pas correct, et nous ne pouvons pas trouver le résultat comme prévu. Mais les recherches que nous avons fait peut faites-leur savoir comment utilise les technique de fouille de donnée dans la domaine de télécommunication.

2 Introduction

Le 3 avril 1973, M. Mation COOPER le directeur général de la division communication de Motorola, à effectuer un appel téléphonique à Joel ENGEL, son rival et néanmoins confrère chez Belle Labs. c'est la premier appel téléphonique en extérieur, L'idée du téléphone portable devient une réalité.

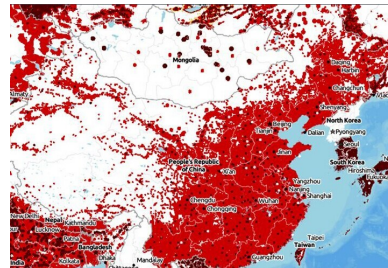
depuis ce jour, le technique développé très rapidement. dans les 20 dernières années, il y a déjà quatre génération des standards pour la téléphonie mobile, non seulement nous pouvons appeler les autres, les nouvelles technologies et les Smart-phones nous permettons aussi envoyer les message, surfer l'Internet, utiliser le service RTSP(Real Timide Streaming Protocol), et le service VoIP (Voice over Internet Protocole),etc.. les services de communication téléphonique sont devenus un outil très important dans notre vie.

2.1 Introduction du CMCC

Fondé en 3 Septembre 1997, après le regroupement de opérateur des télécommunications en 2008, CHINA MOBILE COMMUNICATIONS CORPORATION (CMCC)^{1(a)} est devenu un de trois opérateur des télécommunications en Chine (deux autres sont China Unicom Co., Ltd. et China Telecom). Après plusieurs années de développement, il a construit le plus grand réseau de communications mobiles dans le monde, possède la plus grande base d'utilisateurs dans le monde^{1(b)}. En 2013, le CMCC a 767 million utilisateurs, 630,2 billion ¥ de revenu, 121,7 billions ¥de revenus net, effectif 197,030.



(a) Logo de China Mobile



(b) Réseau télécommunication

FIGURE 1 – CMCC

2.2 La crise du CMCC

Mais en même temps, le taux de croissance des nouveaux utilisateur décline de 22,5 % (2006) à moins de 5% 2013 ². Et dans la premier 3 mois, l'entreprise une fois considérés comme la plus rentable de Chine, le taux de croissance des revenu net est 0,3%.

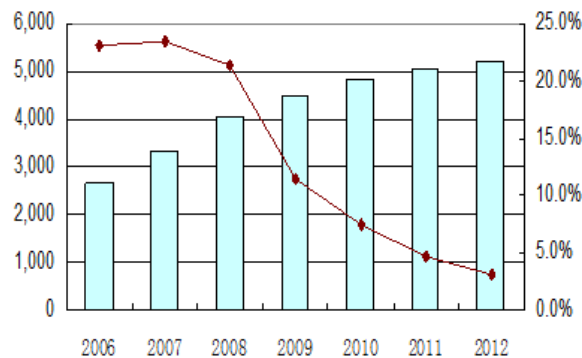
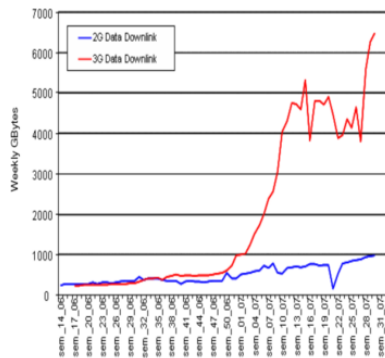
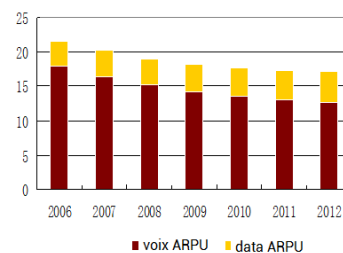


FIGURE 2 – le taux de croissance est décliner

Opérateur des télécommunications Vodafone a fait un étude après il déployé un réseau 3G(the third generation of mobile phone mobile communication technology standards). Comme le réseau 3G permettant des débits (de 2 à 42 Mb/s définis par la dernière génération des réseaux) qui sont bien plus rapides que la génération précédente, par exemple le GSM. Les utilisateur utilisent bien plus souvent le service internet^{3(a)}. Comme ils utilisent plus du service internet, le data ARPU (Average Revenue Per User) augment, mais le voix ARPU décline plus rapide que la montant de data ARPU^{3(b)}.



(a) Downlink Data Traffic in 2G/3G Network



(b) étude de Vodafone

FIGURE 3 – Vodafone

Mais l'étude de Orange nous montre que si nous pouvons fournir des nouveaux technologies qui a plus haute débit, les utilisateur utiliseront plus souvent le service data. ⁴

Traffic per user per technology used

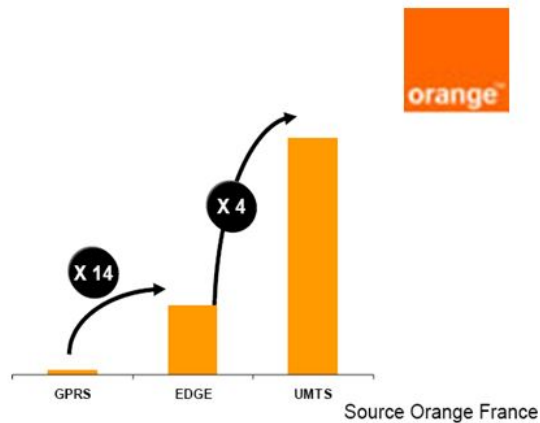


FIGURE 4 – trafic par personne

Des études nous montre nouveaux technologie (comme LTE) peut diminuer le prix de revient, qui peut assurer le profit de l'opérateur. Mais déployer les nouveaux matériel coûte très cher, en 2009, le CMCC dépense 30 milliards ¥ en construit les stations pour réseau 3G, et à 2014, le CMCC a construit 1,5 million stations, à la fin de cette année, il y aura 1,8 million stations, parmi ces stations, il y aura 500 mille stations TD-LTE. En ajoutant des équipements 4G, il peut être mis à niveau une station de 3G à 4G. Donc déployer le réseau 4G n'est pas trop cher, selon l'expérience précédente (de 2G à 3G), les utilisateurs iront utiliser plus le service internet, qui peut assurer le profit de l'entreprise.

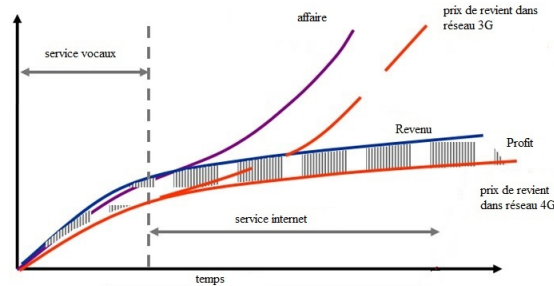


FIGURE 5 – 4G est plus rentable

2.3 L'optimisation du réseau

A part de la évolution des technologies. Un grand enjeu pour les opérateurs est: l'optimisation du réseau télécommunication. Le réseau de communication mobile est très dynamique, la répartition de la densité du trafic est inégale, fréquence très limité, etc. La configuration du réseau état toujours sous-optimal, et la perception de l'utilisateur n'est pas très bien. Donc tous les opérateurs doivent toujours reconfigurer/optimiser/maintien les paramètre du réseau.

Les opérateurs peuvent percevoir les données sur Internet, et utilisent ces informations pour trouver les défauts du système, peut aide l'entreprise optimiser le système.

Mais la optimisation du réseau télécommunication est difficile parce-que: Les technologies d'optimisation de réseau concerné: La technologie de commutation, la technologie sans fil, la configuration et commutation de la fréquence, la signalisation système, l'analyse de trafic, etc. c'est un travail difficile, exiger une meilleure aptitude des employés.

Actuellement, l'optimisation du réseau dépend principalement à la expérience du personnel. Mais des fois les expériences ne sont pas correct. Par exemple, Si l'entreprise besoin de savoir le congestionné d'un station, il faut envoyer les employé avec des équipement pendant les périodes de pointe, mais on ne sait pas si les résultats sont correct [6](#). En outre, souvent un seul type de donnée ont utilise pour l'analyse et la comparaison

pour optimiser les réseau, plutôt que de trouver un solution d'optimisation basées sur toutes les données liées au réseau (telles que les données statistique de trafic, les données d'essai, etc). Et en raison de l'énorme quantité de données, c'est difficile de traite en temps opportun. il est évident que ce méthode est défectueux. Les défauts du système provoque la satisfaction des utilisateurs inférieure, ce qui a conduit à multiplier.

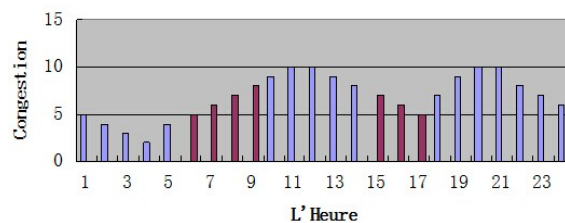


FIGURE 6 – Mesure la congestionné d'un station

Face à des problèmes complexes, les grands entreprises commence utilise les techniques de Fouille de données. Ce technique peut aide l'entreprise faire les décision plus vite et plus précis.

De ce faire, en Juillet 2013, le CMCC a lancé ce projet avec quatre laboratoires dans trois université, ils sont [Tsinghua University](#), [Shandong University](#) et [University of Electronic Science and Technology of China](#). Le projet inclure trois partiel: Fouille de données, gérés le Cloud plateforme et modélisation de l'information dans le système.

2.4 Introduction du laboratoire

De 20 Avril 2014 à 20 Juillet 2014, je fait mon stage chez [laboratoire of Next Generation Network Technology & Application \(NGN\)](#) 7. C'est d'un subordonné de [Research Institute of Network And Human-Machine Speech Communication](#), Département Ingénierie électronique, Tsinghua University. Le laboratoire se trouve dans la ROHM bâtiment.



FIGURE 7 – Logo NGN

Le principaux axes de recherche sont Théorie des réseaux, Architecture de l'Internet, Traitement de l'information Internet, La recherche dans le domaine de la sécurité Internet, Sentiment analyse, Information hiding, etc.

Mon tuteur professionnel est [M. Yongfeng HUANG](#), vice-directeur de la laboratoire NGN. Dans le laboratoire, il y a cinq groupe, chaque groupe a un docteur et son sujet. dans notre groupes, il y a trois personnes, un étudiant de premier année docteur, un étudiant de M1, et une étudiante de Licence troisième année. On utilise R et Rstudio, et Hadoop aussi.

2.5 Objectif du projet

Dans cet article, nous avons d'abord présente le réseau communication mobile, ensuite je vais décrire l'état de l'optimisation du réseau. Enfin je présente la mise en place de notre programme de recherche.

3 Introduction de l'industrie de la télécommunication

3.1 L'évolution des normes de téléphonie mobile

Depuis 1984, il y a déjà plusieurs standards ont été utilisé par les opérateurs dans le monde entier. Voici un tableau de différents standards mobile en Europe et ses paramètres [1](#).

Génération	Acronyme	Description	Débit
1G	Radiocom 2000	Échanges de type voix uniquement	analogique
2G	GSM	Échanges de type voix uniquement	9,05 kbps
2,5G	GPRS	Échange de données sauf voix	171,2 kbps / 50 kbps / 17,9 kbps
3G	UMTS	Voix + données	144 kbps rurale, 384 kbps urbaine, 1,9 Mbps point fixe / -
3.5G ou 3G+ ou H	HSPA	Évolution de l'UMTS	14,4 Mbps / 3,6 Mbps / -
4G	LTE	Long Term Evolution (Données)	150 Mbps / 40 Mbps / -
4G	LTE-Advanced	Long Term Evolution Advanced (Données+voix)	1 Gbps à l'arrêt, 100 Mbps en mouvement / - / -

TABLE 1 – Les différentes générations de téléphonie mobile en Europe

3.1.1 La premier génération

En télécommunication, 1G est la premier génération des standards pour la téléphonie mobile, il s'agit de la première apparition du réseaux de téléphonie mobile, 1G sont des réseaux analogiques, peut échanges de type voix uniquement.

3.1.2 La deuxième génération

2G, la technologie de téléphonie sans fil de deuxième génération, la différence entre le réseaux 1G et 2G est: le signaux radio sur les réseaux 1G sont analogiques, et celle de 2G sont numériques.

Systèmes 2G ont été significativement plus efficaces du spectre permettant de bien plus grand taux de pénétration du téléphone mobile, en plus les données vocales numériques peuvent être compressées et multiplexées beaucoup plus efficacement que les codages de la voix analogique grâce à l'utilisation de codecs différents, ce qui permet plus d'appels à transmettre dans la même quantité de bande passante radio. Et 2G présenté premier foi le service de données pour mobile. Les Technologie 2G permettent les divers réseaux de téléphonie mobile de utiliser des services tels que le SMS et MMS. Tous les message de texte envoyés au delà de 2G sont chiffrés numériquement, ce qui permet le transfert de données de telle sorte que seul le destinataire peut recevoir et lire.

Réseaux 2G ont été construits principalement pour le service téléphoniques et de transmission de données lent (défini dans les documents de spécifications IMT-2000).

Réseaux 2,5G, on le qualifie souvent de le General packet Radio Service ou GPRS, est une norme pour la téléphonie mobile dérivée du GSM et complémentaire de celui-ci, permettant un débit de données plus élevé. Le 2,5 indique que c'est une technologie à mi-chemin entre le GSM (deuxième génération) et l'UMTS (troisième génération). Le GPRS est une extension du protocole GSM : il ajoute par rapport à ce dernier la transmission par paquets. Cette méthode est plus adaptée à la transmission des données. En effet, les ressources ne sont allouées que lorsque des données sont échangées, contrairement

au mode « circuit » en GSM où un circuit est établi - et les ressources associées - pour toute la durée de la communication. Le GPRS a ensuite évolué au début des années 2000 vers la norme EDGE également optimisée pour transférer des données et qui utilise les mêmes antennes et les mêmes fréquences radio.

3.1.3 La troisième génération

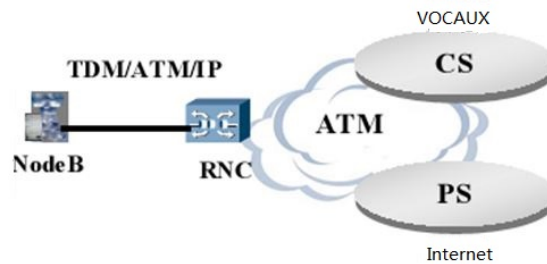
La troisième génération (3G) des normes de téléphonie mobile. Elle est représentée principalement par W-CDMA, CDMA2000, TD-SCDMA et WiMAX. Elle permettant des débits de 2 à 42 Mb/s qui sont bien plus rapides qu'avec la génération précédente. Grâce à l'utilisation des règles de classement de l'utilisateur, et les bandes de fréquences supérieures rendant la capacité du réseau augmenter.

Dans les différents standards 3G et ses prédécesseurs, ils utilisent le domaine CS (Circuit Switch) pour le service vocal, et le domaine PS (Packet Switch) s'occupe du service de données 8(a).

3.1.4 La quatrième génération

La quatrième génération des standards pour la téléphonie mobile, succédant à la 2G et la 3G, en théorie, elle permet de transmettre de données à des débits supérieurs à 100 Mb/s.

Une des particularités de la 4G est sa EPC (Evolved Packet Core) est basé sur IP, et il n'y a plus de mode commuté (le 'Circuit Switched Domain' qui s'occupe le service vocal dans les standards précédents), ce qui signifie que le service vocal est transmis sur l'internet 8(b).



(a) Réseau 3G et ses prédécesseur



(b) Réseau 4G

FIGURE 8 – Structure des réseaux

Les avantages du réseau 4G sont: plus haut débit, mieux utilisation de la bande de fréquence, moins de délai (délai dans le panneau de l'utilisateur est inférieur que 5 ms, délai dans le panneau de commande est inférieur que 100 ms), plus simple structure du réseau, moins de consommation d'énergie Terminal.

3.2 Le réseau LTE

Le LTE (Long Term Evolution) est l'évolution la plus récente des normes de CDMA 2000, TD-SCDMA, GSM. La norme LTE. La technologie LTE a été considérée comme une norme de troisième génération '3.9G', et la 'vraie 4G', appelée LTE-Advanced a été reconnue par l'UIT comme une technologie 4G en 2010. LTE a deux branches: LTE-FDD (Frequency-Division Duplex Long Term Evolution) et LTE-TDD, (Time Division Duplex Long Term Evolution) les deux standards sont similaires, la différence entre les deux standards est moins de 15% ⁹. En 2011-2012, les

réseaux LTE-TDD sont commercialisés sous l'appellation 4G par le CMCC un Chine.



FIGURE 9 – l'évolution des standard

3.2.1 La structure du réseau LTE

Le réseau 4G contient 3 partie: UE(User Equipment);, eNodeB (les stations de base), EPC (Evolved Packet Core). EPC contient MME, S-GW, P-GW et HSS [10 2](#).

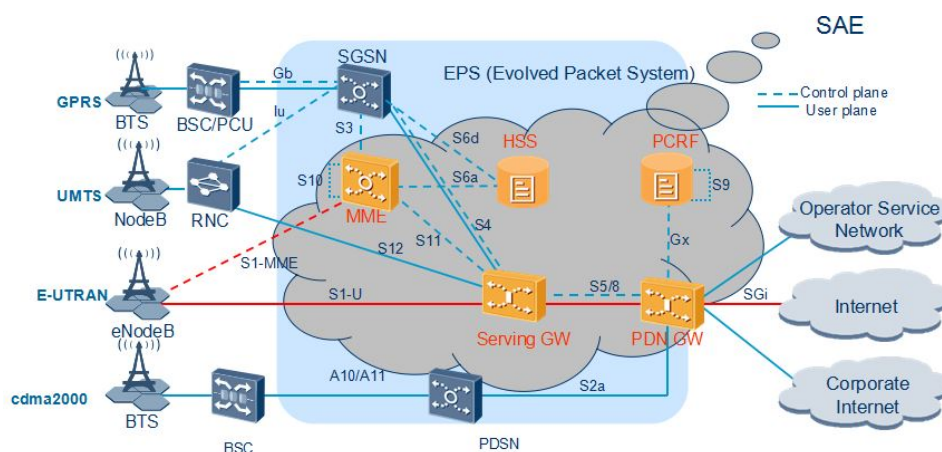


FIGURE 10 – la structure du réseau

Part	Fonction
MME	L'authentification des utilisateurs et la gestion des clés, Cryptage de la couche NAS, Gestion de la liste TA, Sélection P-GW ou S-GW
Service Gateway	Compression d'en-tête IP, Routage de paquets et la transmission, La commutation entre eNB, Facturation des utilisateurs porteur
PDN Gateway	L'allocation des adresses IP de UE, l'accès aux fonction de gestion de réseau externes, Facturation en service
HSS(Home Subscriber Service)	Stockée données de l'utilisateur associées au service
PCRF	Roaming

TABLE 2 – la fonction du chaque partie

Entre deux E-UTRAN, il y a l'interface X2, l'interface S-11 se trouve entre S-GW et MME, E-UTRAN et S-GW échange les données par l'interface S1-U et il échange les donnée par l'interface S1-AP avec MME, MME et HSS utilise l'interface S6A, et l'interface S5/8 entre S-GW et P-GW, Gx entre PCRF et P-GW. En mettant des capteur en les interfaces, les opérateurs et les fournisseurs d'équipement peuvent collecter les données de signalisation, et utilisent ces informations pour trouver les défauts du système.

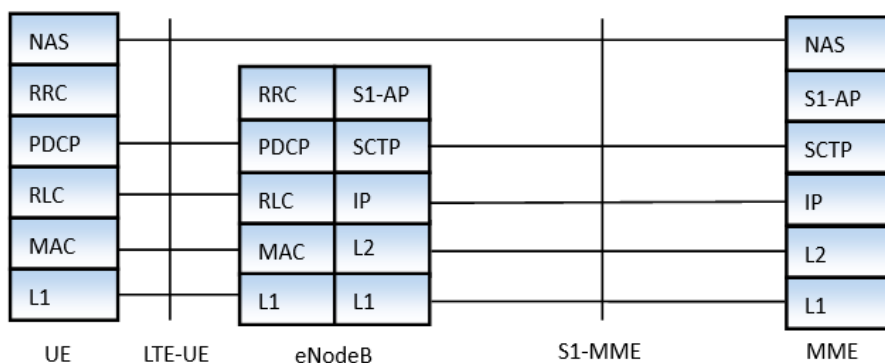


FIGURE 11 – Contrôle plan

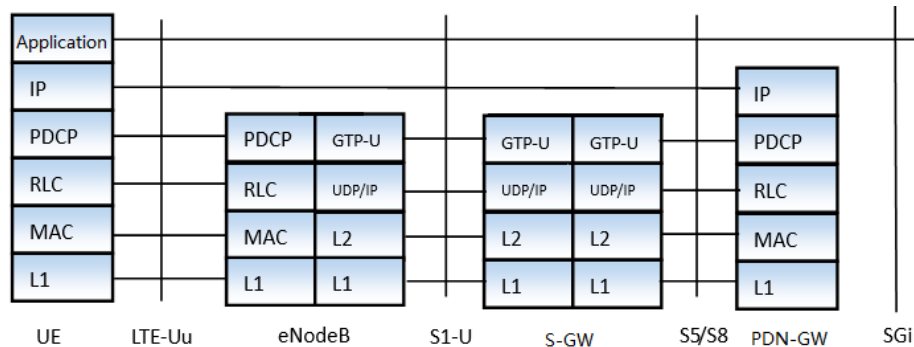


FIGURE 12 – User plan

4 Les solution existant

L'optimisation du service téléphonie est très important. L'opérateur a construit un immense réseau télécommunication, mais a cause de la mauvaise configuration du système, les utilisateur ne sont pas satisfaits avec les services, les investissement n'a pas été remis. Donc les entreprises comme IBM, Huawei, et l'autre fournisseur du équipement essaient de trouver la meilleure solution.

Maintenant, il y a beaucoup des gens travaillent sur ce domaine, nous avons trouvé beaucoup de papier sur l'optimisation du réseau télécommunication, mais les articles sont basé sur réseau 3G ou 2G.

Il y a trois techniques qui sont beaucoup utilisé:

1. la Technique KQI;
2. la Technique QoE;
3. la Technique qui étudie les comportements de l'utilisateur.

la Technique KQI:

La technique utilise le plus souvent s'appelle 'KQI' (Key Quality Indicator) [13](#), cette méthode a été beaucoup utilisé. Et cette technique peut généralement divisé en deux étapes. d'abord, nous devons calculer le score de KPI, pour calcule le KPI en premier, il faut analyser le processus d'un service et choisir les indicateur de performances. Ensuite, nous pouvons calculer le score d'un processus en utilisant un

équation linéaire, le poids de chaque attributs change selon le service, par exemple, pour le service SMS, le délai porte peu de l'importance, mais le délai du service est important pour le service HTTP. à la fin, nous pouvons calculer le KQI à avec les KPI [2]. Mais les poids sont défini par les experts, et les valeurs peut-être fausse ou pas précis. Et par fois le score est bonne mais l'expérience de l'utilisateur n'est pas bon.

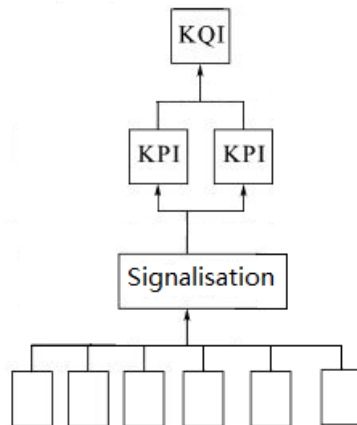


FIGURE 13 – KQI

la Technique QoE:

KPI est un des indicateurs de qualité axés sur les performances du réseau, mais il ne reflète pas directement l'expérience de la qualité de service de l'utilisateur, parce que les expériences de l'utilisateur sont difficile à mesure. Donc la technique QoE a été inventé. QoE défini la performance, de la qualité de service et l'expérience de l'utilisateur de l'ensemble du réseau à partir de l'utilisateur.

Les utilisateurs ont nombreuses exigences pour les services téléphonie, ils peuvent être résumées comme deux aspects: la fiabilité et le confort. La fiabilité fait référence à l'activité de l'accessibilité, la disponibilité et la durabilité. Le confort est une qualité de service, est un indice de la perception directe de l'utilisateur, qui dépend à l'expérience de l'utilisateur[3]. Les relations entre QoE et QoS KPI sont: la fiabilité du service 3, le confort du service4.

TABLE 3 – Fiabilité du service

KQI	QoE
Accessibilité	Taux de succès
Disponibilité	Temps d'accès aux services
Durabilité	La durée de l'accès des services

TABLE 4 – Confort du service

KQI	QoE
	Taux de perte de paquets de couche d'application
	Le débit moyen
La qualité du service de transmission	Stabilité de la transmission
	Le bout en bout délai moyen
	Gigue
Le persistant de la connexion de service	La vitesse et la difficulté du service d'assistance

Maintenant, la technique de QoE a été beaucoup utilisé pour le service vocaux. Et grâce à la complexité du service de données, il n'y a pas un standard de QoE pour le service de données.

En utilisant la technique KQI et QoE, nous peuvent mesurer la qualité du service, les résultats peuvent aider les opérateurs trouver les services de mauvaise qualité, les opérateur peuvent améliorer les services selon le résultat, finalement améliorer la notation de l'utilisateur.

Le résultat de KQI dépend seulement aux performances du réseau, donc nous avons besoin les informations des performances de réseau. Et la technologie QoE besoin Le résultat de KQI et les feed-back de l'utilisateur, le feed-back

peut obtenir par l'enquête ou les plaintes des utilisateurs, et par les mesures directs.

la Technique qui étudie les comportements de l'utilisateur

Aussi il y a un groupe qui utilise le comportement de l'utilisateur pour définir la qualité du service[4]. Le groupe utilise cette méthode dans le service vocaux, il cherche la situation comme l'utilisateur accroche et ré-appelle la même personne. À la fin, cette méthode aide l'opérateur à corriger le paramètre d'erreur.

Selon l'article, cette méthode peut aider l'opérateur à trouver les défauts du système, mais il y a de nombreuses restrictions, par exemple, nous ne pouvons pas utiliser cette technologie dans le service de SMS, etc.

5 La présentation de notre solution

La méthode qui utilise le comportement de l'utilisateur est intéressante, mais nous trouvons qu'elle peut être utilisée seulement dans le service vocaux, nous n'avons pas trouvé des règles similaires dans l'autre service. D'ailleurs, le réseau LTE ne supporte pas le service vocaux, donc il n'existe pas d'optimisation du service vocaux dans le réseau LTE et nous n'avons pas de données. Donc nous ne pouvons pas utiliser cette méthode.

La technique QoE et KQI sont beaucoup utilisées, mais d'abord, pour la méthode QoE, nous avons besoin des réponses des utilisateurs, mais nous n'avons pas assez de temps, et des raisons financières, le CMCC ne peut pas nous fournir ces données. Et l'équation qu'on utilise pour calculer KPI n'est pas convaincante, voici un exemple d'une équation pour calculer la disponibilité du réseau pour le service SMS dans le réseau 3G[14].

$$\begin{aligned} \text{RANQuality} &= \text{PDCH_AllocationFailure} * 25\% \\ &+ \text{PRACH_ImmediateAssignRejection} * 25\% \\ &+ \text{PRACH_ReqFailure} * 20\% \\ &+ \text{Prob_RLC_Retransmission} * 15\% \\ &+ \text{RadioChannel_BLER} * 15\% \end{aligned}$$

FIGURE 14 – Un exemple d’une équation pour calculer la disponibilité

Le poids de chaque attribut est défini par les experts, Mais l'utilisateur n'est pas satisfait du service. Nous croyons que l'erreur a été causée par l'inexacte équation, et nous pensons que les algorithmes de Classification peuvent aider à améliorer le résultat, mais très vite nous avons trouvé que le CMCC ne peut pas nous fournir ce type de données. Sans la connaissance a priori, nous ne pouvons pas utiliser ces algorithmes. Nous avons aussi pensé à utiliser l'externalisation ouverte (crowdsourcing), à notre avis si le CMCC peut lancer un projet d'externalisation ouverte, si le CMCC peut encourager ses utilisateurs donnent les notes aux services pour obtenir des crédits, nous pouvons obtenir la connaissance a priori, et à l'aide de ces données, nous pouvons trouver une équation peut-être mieux que les équations écrites par les experts. Mais bien sûr, le CMCC n'a pas accepté cette idée, parce que cette méthode peut coûter cher, et peut-être il n'y a pas de revenu direct. Et l'entreprise ne fait pas de l'investissement sans retour.

Finalement, nous avons décidé d'utiliser l'algorithme de clustering. C'est une méthode statistique d'analyse des données. Elle divise un ensemble de données en différents groupes, les données de chaque groupe ont mathématiquement plus de proximité que les données de l'autre groupe, et nous supposons que les données dans la même partition ont des caractéristiques similaires.

5.1 Clustering

L'algorithme de clustering est une des méthodes de classification non supervisée. Il est beaucoup utilisé quand le donnée n'a pas de connaissance a priori.

Il existe de multiples méthodes de regroupement des données, parmi lesquelles :

- Classification basées sur la densité;
- Classification hiérarchique;
- Classification par partitionnement;
- Classification par grille;
- Classification basées sur des modèles.

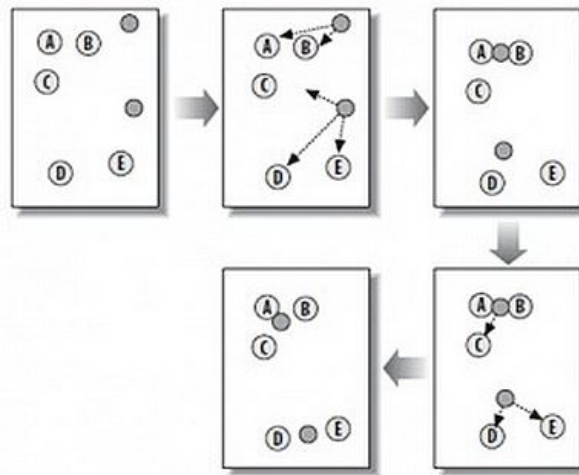


FIGURE 15 – procédé de l'algorithme K-Means

Nous avons utiliser l'algorithme K-Means et CLARA (Clustering Large Applications). Ils sont deux techniques de Classification par partitionnement.

Le but de cet algorithme est de diviser des données en K partitions (clusters) dans lesquelles les données appartient à la partition avec la moyenne la plus proche.

5.1.1 La distance

Quand on regroupe un ensemble de données, on calcule la distance entre chaque échantillon, la distance entre les

échantillons dans un même groupe sont plus petit que les échantillons dans l'autre groupe. on suppose que les plus semblables les deux échantillons, plus la distance est petit, plus la similitude est grand.

Il y a plusieurs moyen de calculer la similarité, On utilise la technique qui basé sur la distance, les techniques utilisé le plus souvent sont: *distance euclidienne*, *distance de Hamming*, *distance Manhattan*, etc. Aussi il existe des méthodes qui mesuré la similarité, il inclut: *coefficient de Jaccard*, *cosinus similarité*, etc . Nous décidons de utiliser la distance euclidienne pour calculer la similarité.

$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

n est la dimension, x_k et y_k sont le k -ème attributs de échantillon x et y .

5.1.2 La validité de clustering

C'est méthode a été utilisés dans beaucoup différents domaines. Mais la qualité du résultat dépend au nombre de clusters ou la valeur de K , différent paramètre peuvent mener à des résultat très différent.

En 1974, Mme BEZDEK a proposé cette question de évaluation, et donne la première fonction pour évaluer le résultat du clustering: Partition Coefficient. Par la suite de nombreux chercheurs ont proposé une variété de fonctions pour évaluer le résultat, et évalue les qualité et le champ d'application de ses fonctions.

Par mesure l'efficacité entre les partitions et inter-partition, on peut évaluer le clustering. Le résultat de clustering idéal devrait être avoir la distance minimale dans le partitions, et la distance maximale entre les partitions. C'est à dire avec le plus petit grand de la cohésion dans la partition, et le plus élevé degré de la séparation entre les partitions. La cohésion mesure le degré de proximité dans la

groupe, et la séparation mesure le degré de dissimilarité entre les groupes. La cohésion et la séparation peuvent être calculées par les équations suivantes.

$$cohésion(C_i) = \sum_{x \in C_i} proximity(x, c_i)$$

$$séparation(C_i, C_j) = proximity(c_i, c_j)$$

Dans les formules, le c est le centre de gravité d'une partition, $proximity(a, b)$ est la mesure de proximité entre la partition a et b .

Le degré de la cohésion de clusters et de la séparation est souvent utilisé comme une mesure principale pour évaluer le résultat de clustering. Un bon regroupement devrait être à la fois un petit degré de la cohésion et un grand degré de la séparation.

Le degré de la cohésion et de la séparation de la partition ne sont pas indépendants, la somme des deux est une constante, on suppose que quand on a le minimal degré de la cohésion, on peut avoir le maximal degré de la séparation. Évidemment, nous devons utiliser à la fois le degré de la cohésion et de la séparation pour mesurer la qualité du regroupement.

5.1.3 La silhouette Coefficient

Kaufman a proposé la silhouette coefficient en 1979, cette méthode utilise à la fois les deux degrés.

1. la silhouette coefficient d'un échantillon:

Pour un échantillon d_i , en supposant qu'il est dans le groupe A , la silhouette coefficient peut être calculée par la formule:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

a_i est la dissimilitude moyenne de la donnée i avec toutes les autres données dans le même cluster (plus la valeur est petite, meilleur est le regroupement). b_i est la différence moyenne de i à tout autre groupe i n'est pas un membre.

Qui peut être écrite comme :

$$s_i = \begin{cases} 1 - \frac{a_i}{b_i}, & \text{if } a_i < b_i \\ 0, & \text{if } a_i = b_i \\ \frac{b_i}{a_i} - 1, & \text{if } a_i > b_i \end{cases}$$

D'après la définition ci-dessus, il est clair que la valeur de s_i est entre 1 et -1

$$-1 < s_i < 1$$

Pour s_i d'être proche de 1, nous demandons le a_i plus petit que b_i . Comme a_i mesure la dissemblance de i et son propre groupe, une petite valeur signifie qu'il est bien adapté. En outre, un grand b_i implique que i est mal adapté à son groupe voisin. Ainsi, un s_i près de 1 signifie que la donnée est concentrée de manière appropriée. Si la valeur de s_i est proche de -1 , selon la même logique, nous voyons que la donnée i serait plus appropriée si elle a été regroupée dans son groupe voisin.

2. la silhouette coefficient moyenne :

Pour le résultat d'un regroupement, le silhouette coefficient égale à :

$$s_k = \frac{1}{n} \sum_{i=1}^n s_i$$

n est le nombre de données, k est le nombre de partition, s_k est la moyenne du silhouette coefficient. Nous pouvons utiliser s_k pour évaluer la qualité du clustering.

Le s_i est les moyennes sur l'ensemble des données d'un cluster, il est une mesure du degré de cohésion de toutes les données du cluster. Ainsi, les moyennes de s_i de l'ensemble des données (s_k) est une mesure de la façon appropriée les données ont été regroupées. S'il y a trop peu de clusters, ainsi que peut se produire lorsque un mauvais choix de k est utilisé dans l'algorithme K-Means, la silhouette Coefficient de certains groupes est beaucoup plus petit que les autres. Donc la plot de silhouette Coefficient et la valeur moyenne de silhouette peuvent être utilisés pour déterminer le nombre de cluster (la valeur de k) optimal pour l'ensemble de données.

Après trouver les paramètre optimal, nous pouvons utiliser l'algorithme clustering, et après étudié et comparé les caractéristiques de chaque partition, on pouvons défini quelle partition représentent mauvais qualité de service. Les résultat peut nous aider à classifier le service.

5.2 Règles d'association

La règle d'association est une méthode populaire, elle étudiée d'une manière approfondie dont le but est de découvrir des relations ayant un intérêt pour le statisticien entre les variables. En se basant sur le concept de relations fortes, Rakesh Agrawal et son équipe présente des règles d'association dont le but est de découvrir des similitudes entre des produits dans des données saisies sur une grande échelle dans les systèmes informatiques des points de ventes des chaînes de supermarchés. Par exemple, une règle découverte que si un homme achète les serviettes de bébé, il est susceptible de achète les bières. Une telle information peut être utile quand on veut prendre des décisions marketing.

Les règles d'association sont employées aujourd'hui dans plusieurs domaines, incluant: la fouille du web, la détection d'intrusion et la bio-informatique. Dans ces domaines, ils utilisent les données booléenne pour trouver les règles utiles. Mais dans le domaine télécommunication, les données de signalisation sont le données numérique. Donc nous ne pouvons pas directement utiliser la règles d'association. Par contre nous pouvons convertir les données de type numérique en données de caractère en divisant les données dans plusieurs partitions [6]. Par exemple, nous avons des données numérique entre 0 et 100, et on divise les données en 10 partitions, donc les chiffres entre 0 et 10 peuvent présenter par $0 < x < 10$. En utilisant cette technique nous pouvons utiliser la règle d'association pour trouver les relations dans les données.

Il y a plusieurs méthodes pour diviser les attributs numériques: par catégorisation, l'analyse typologique, par analyse de l'histogramme, l'analyse basé sur l'entropie de discret, partition naturel et ainsi de suite.

Nous décidons de utiliser le K-Means d'abord. Après analyse

le résultat, nous pouvons utiliser la algorithmes règles d'associations à l'aide du résultat de K-Means.

5.3 La mise en œuvre

5.3.1 Le logiciel utilisés

L'objectif de notre projet est de trouver une méthode qui peut s'implanter dans les serveurs du CMCC, et aider le CMCC améliorer la qualité du réseau. D'après le directeur du R&D département du CMCC, au total, il a 20Pb de donnée stock dans ses base de données, donc le logiciel doit être capable d'exécuter grand quantité de données. En plus, au lieu de utilisé directement les logiciel comme SQL (qui n'est pas très efficient si on a beaucoup de donnée stock dans différent serveur), le CMCC utilise 'Hadoop' pour stock et géré ses données.

Donc le logiciel que nous utilisons doit être capable d'exécuter grand quantité de donnée et peut travailler avec Hadoop.

Finalement nous décidons de utiliser le langage R.
L'avantage de R sont:

- R est un langage et un environnement pour le calculé statique et les graphiques;
- R offre une grande variété de statistiques (modélisation linéaire et non linéaire, classification, clustering, etc,) et des graphiques techniques, et il est très extensible;
- R est facile à utiliser;
- R est un logiciel libre, il compile et fonctionne sur une grande variété de plates-formes UNIX et les systèmes similaires (y compris FreeBSD et Linux), Windows et Mac OS;
- en utilisant les packages fournir par 'Revolution Analytics', nous pouvons utiliser Hadoop en R.

Nous utilisons Rstudio comme notre environnement de programmation. C'est une interface utilisateur puissante et productive pour R.

5.3.2 Introduction des données

Après quelque semaines de négociation avec les employés de différents départements de le CMCC, ils nous ont fourni deux versions de données, et ses spécifications du format[1]. Nous avons trouvé que le CMCC n'a pas de accès direct aux données, et le fournisseur d'équipement a modifié le spécification fournir par le CMCC, et il y a des erreurs dans les données fourni par les fournisseur d'équipement.

Ils nous ont envoyé 11 dossiers, chaque dossier correspond à un service. les services sont 'rtsp', 'dns', 'mail', 'ftp', 'http-wap', 'mms', 'p2p', 'realtimecom', 'VoIP' et les données de signalisation entre E-UTRAN et MME 'S1AP-NAS'.

Et nous avons trouvé que pour les services comme 'VoIP' et 'RTSP', ils sont très peu de données 5. Donc nous avons décidé de utiliser le donnée du service 'HTTP'.

L'interface	Nombre de ligne
S1-AP	240
RTSP	35
DNS	272562
Mail	44
FTP	71
HTTP-WAP	50854
MMS	193
P2P	515
Realtimecom	2082
S1U	89759
VoIP	28

TABLE 5 – les dossiers de données

Le dossier du service HTTP a 18,4Mbit , il y a 50854 lignes, tous les données sont collectées par les capteurs placer entre les Service-Gateway et les eNodeB. le capteur enregistre un ligne de donnée quand un processus est fini. chaque ligne a 76 attributs16.


 data.http	50894 obs. of 76 variables
---	----------------------------

FIGURE 16 – les données du service HTTP

Il contient des informations de UE (IMEI, IMSI, etc), les trafic de la liaison montante et la liaison descendante et le temps, la adresse IP de UE, eNodeB et S-GW, le port de UE, eNodeB et S-GW, le délai du service, le site web, cookie, et aussi le début temps et le temps d'arrêter. les données sont collectent dans 20.92 minutes 17.

```
StartTimes <-as.POSIXlt(as.numeric(substr(data_HTTP$StartT,1,10)), "UTC", origin="1970-01-01")
EndTimes   <-as.POSIXlt(as.numeric(substr(data_HTTP$StopT,1,10)), "UTC", origin="1970-01-01")
RecodeTime <-data.frame(StartTimes,EndTimes)
attach(RecodeTime)
```

```
cat("Data recoded in:",max(EndTimes)-min(StartTimes)," minits")
```

```
## Data recoded in: 20.92 minits
```

FIGURE 17 – Les données sont collectent dans 20.92 minutes

5.3.3 Prétraitement de données

En analysant des données, nous avons trouvé des erreurs de données, et le fournisseur nous a confirmé que ces sont les défaut de leur système 4G. Pour les attributs 'IMSI', 'IMEI', 'MSISDN', 90 % de lignes sont vides, ce qui ne sont pas vides, les contenus sont illisible, et peuvent provoquant des erreurs de lecture 18. Et nous avons trouvé que dans les contenus de certains lignes sont bizarre.

IMSI	IMEI	MSISDN	M.TMSI	IpType
			o 1	175453935
hw2		hQp p	1	176899993
			H 1	176918718

FIGURE 18 – erreur du codage BCD

Dans ce processus, le 'Down Link Online Time' égale à 0 ms , mais il a téléchargé 746 bits, c'est clairement un erreur du données.

BigType	SubType	L4	ServerIP	ServerPort	UpTraffic	DownTraffic	UpTime	DownTime
15	5017	0	3719544451	80	595	746	500	0

FIGURE 19 – Erreur de la donnée

Nous avons décidé de ne utiliser les données avec ce type de erreur, à la fin, en supprimant ses données, il nous reste 37865 lignes (50894 lignes en origine, 13029 lignes ont été supprimé) [20](#).

data_DisHttp	50894 obs. of 76 variables
data_HTTP	37865 obs. of 76 variables

FIGURE 20 – Prétraitement des données

Entre ces 76 attributs, Une grande partie de ces informations sont inutile, et pour certain attributs les contenus égal tous à 0. Finalement nous avons trouvé 11 attributs. ils sont:

Signalisation	Signalisation	KPI
trafic en liaison montante	le temps en ligne	vitesse
trafic en liaison descendante	le temps en ligne	vitesse
	Http First Response Time	délai
	Http Last Packet Time	délai
	Http Last Ack Time	délai
Packet Num en liaison montante	retransmission de paquets Num en liaison montante	taux de retransmission
Packet Num en liaison descendante	retransmission de paquets Num en liaison descendante	taux de retransmission

Mais nous avons trouvé que dans certains lignes le taux de retransmission sont trop grands (plus grand que 100%). par exemple, dans un processus, il a téléchargé 17 paquets IP, et il a 7448 paquets sont désordre, et ré-téléchargé 6384 paquets [21](#). Les données ne sont pas correct, donc nous ne pouvons pas utiliser ses donnée pour calculer le taux de retransmission.

UpPac	DownPac	UpDisPac	DownDisPac	UpRePac	DownRePac
6	5	0	0	0	0
10	19	0	0	0	0
5	4	380	0	0	0
8	17	0	7448	40	6384
3	6	0	2128	0	0

FIGURE 21 – Défaut de la système

Finalement nous avons décidé de utiliser ces 5 attributs (la vitesse et le délai) pour mesurer la qualité du service.

5.3.4 Les caractéristique du donnée

```
> min(upAvBand)      > max(upAvBand)      > sd(upAvBand)
[1] 0.06616667        [1] 170.16              [1] 1.631399
> min(downAvBand)    > max(downAvBand)    > sd(downAvBand)
[1] 0                  [1] 1316                [1] 24.36523
>
> min(firstRespondTime) > max(firstRespondTime) > sd(firstRespondTime)
[1] 0                  [1] 150695              [1] 4475.789
> min(lastPacketTime)  > max(lastPacketTime)  > sd(lastPacketTime)
[1] 0                  [1] 92081               [1] 2673.268
> min(lastAckTime)     > max(lastAckTime)     > sd(lastAckTime)
[1] 0                  [1] 1220150            [1] 8989.971
```

FIGURE 22 – La valeur maximal, valeur minimal et l'écart type

```
> mean(upAvBand)
[1] 0.85507
> mean(downAvBand)
[1] 4.735396
>
> mean(firstRespondTime)
[1] 594.1533
> mean(lastPacketTime)
[1] 392.4637
> mean(lastAckTime)
[1] 1231.584
```

FIGURE 23 – La valeur moyenne

Nous avons trouvé que l'écart type pour les trois attributs de délais sont très grand. Les valeurs varient de $0ms$ à $1220150ms$. Par contre la changement de la vitesse de la liaison montante et descendante n'est pas très grand.

6 La mise en œuvre

6.1 La valeur optimal de k

Pour trouvé le k optimal pour notre données, nous avons utilisé la technique 'silhouette Coefficient' et 'somme de erreur carrés'.

Nous avons mesuré le résultat de ces deux techniques quand la valeur de k augmente de 1 à 15, étant donné que les caractéristique de l'algorithme, pour chaque valeur de k , nous avons répétées 50 fois et calcule la valeur moyenne pour éliminer les erreurs, en suite, nous étudions le résultat pour trouver la valeur optimal.

6.1.1 La somme d'erreur carrés

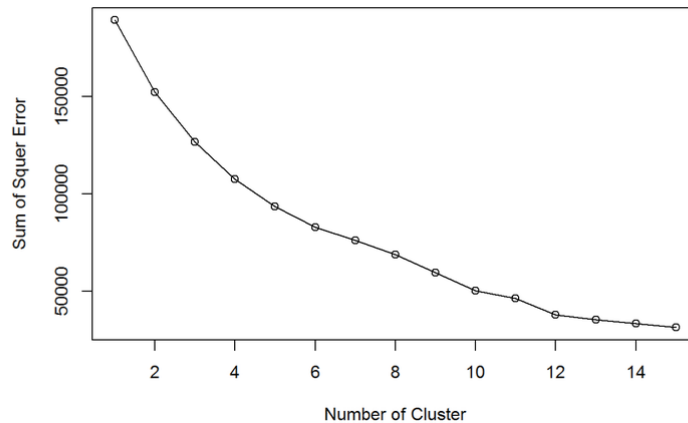


FIGURE 24 – somme d'erreur carrés

Nous avons trouvé que la somme d'erreur carrés diminue quand k augment, mais le taux de déclin est très faible, et il a peu de changements. Donc nous ne pouvons pas utilise cette méthode pour trouver le k .

6.1.2 La silhouette Coefficient

Nous utilisons la même technique pour calculer et visualiser la valeur de la silhouette Coefficient avec la différence valeur de k .

D'abord, nous avons calculé la silhouette Coefficient quand la valeur de k change de 2 à 40:

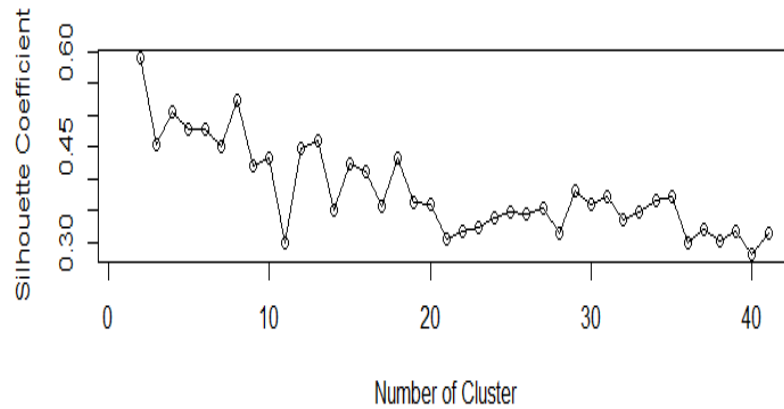


FIGURE 25 – la silhouette Coefficient quand k varie de 2 à 40

Nous avons trouvé que la valeur de la silhouette Coefficient décline quand k augmente. Et les valeurs de silhouette Coefficient de 2 à 15 sont beaucoup plus grand que celle de 16 à 40. Et nous regardons la silhouette Coefficient quand la valeur de k varie de 2 à 15.

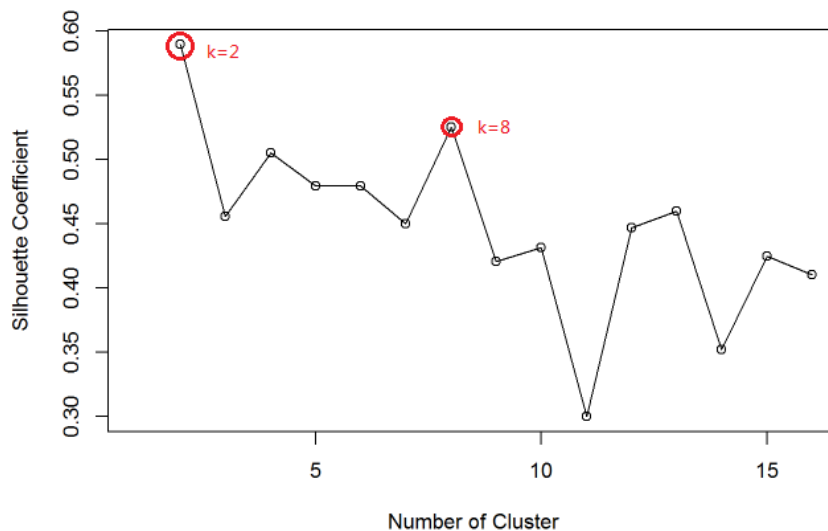


FIGURE 26 – la silhouette Coefficient quand k varie de 2 à 15

Mais l'image de silhouette Coefficient

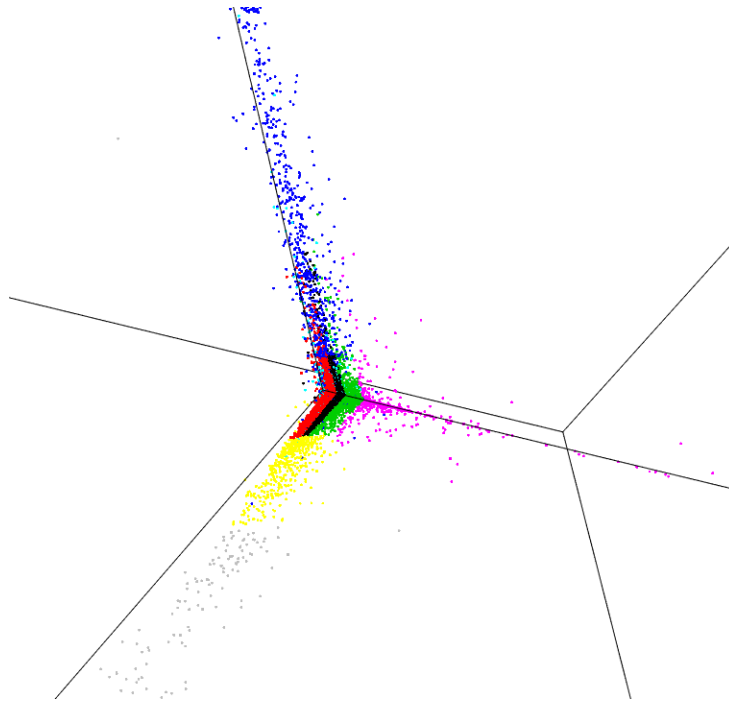


FIGURE 27 –

row.names	V1	V2	V3	V4	V5
minV	0.65666667	0.04000	0	0	0
maxV	1.37400000	36.12733	10944	4001	16132
minV	0.06616667	0.00000	0	0	0
maxV	0.84000000	39.52800	10017	3905	15960
minV	1.26400000	0.08000	0	0	0
maxV	2.83400000	34.86400	14231	4256	17640
minV	0.13670588	0.08000	0	4	0
maxV	9.01800000	94.32622	150695	92081	97989
minV	0.10936691	0.12900	0	3	16059
maxV	3.99800000	53.73352	54739	60889	1220150
minV	2.83400000	0.08000	0	0	0
maxV	170.16000000	43.89892	11737	7855	55655
minV	0.21200000	32.65100	0	0	0
maxV	6.19200000	112.48000	3790	2795	10485
minV	0.85011494	89.10400	0	0	0
maxV	15.85081818	1316.00000	37401	1405	32855

FIGURE 28 –

Rapport de Stage

V1	V2	V3	V4	V5
0.08776679	0.02000	1895.0	1.5	5242.5
0.12303639	0.06000	6903.5	3.5	13222.5
0.17435294	0.08000	10480.5	704.5	16009.5
0.43433333	0.08000	11340.5	2100.0	16095.5
0.74833333	0.10450	12984.0	3350.0	16886.0
0.84505747	16.39000	25816.0	3953.0	25247.5
1.05705747	33.75750	46070.0	4128.5	44255.0
1.31900000	35.49567	102717.0	6055.5	76822.0
2.10400000	37.82767	0.0	34372.0	659069.5
2.83400000	41.71346	0.0	76485.0	0.0
3.41600000	48.81622	0.0	0.0	0.0
5.09500000	71.41876	0.0	0.0	0.0
7.60500000	91.71511	0.0	0.0	0.0
12.43440909	103.40311	0.0	0.0	0.0
93.00540909	714.24000	0.0	0.0	0.0

FIGURE 29 –

row.names	upAvBand	downAvBand	firstRespondTime	lastPacketTime	lastAckTime
1	8	6	2	3	1
2	5	7	2	3	8
3	5	10	1	1	1
4	2	12	1	1	1
5	4	7	1	1	1
6	4	12	2	2	1
7	4	7	2	3	1
8	7	6	1	1	1
10	4	6	1	1	1
11	5	6	2	3	1
12	4	6	1	1	1
13	4	10	1	2	1
14	9	6	2	3	1
15	4	12	1	2	1
16	4	7	1	1	1
18	4	12	1	1	1

FIGURE 30 –

Rapport de Stage

```
inspect(aRule)
```

##	lhs	rhs	support	confidence	lift
## 1	{}	=> {lastAckTime=1}	0.9222	0.9222	1.0000
## 2	{lastPacketTime=3}	=> {firstRespondTime=2}	0.1133	0.8438	5.0682
## 3	{downAvBand=7}	=> {firstRespondTime=1}	0.1422	0.8243	1.0380
## 4	{downAvBand=7}	=> {lastAckTime=1}	0.1611	0.9337	1.0125
## 5	{upAvBand=5}	=> {lastAckTime=1}	0.1701	0.8911	0.9663
## 6	{upAvBand=4}	=> {lastPacketTime=1}	0.3021	0.8695	1.2254
## 7	{upAvBand=4}	=> {firstRespondTime=1}	0.3187	0.9174	1.1552
## 8	{upAvBand=4}	=> {lastAckTime=1}	0.3359	0.9668	1.0484
## 9	{downAvBand=6}	=> {lastAckTime=1}	0.5261	0.9041	0.9804
## 10	{lastPacketTime=1}	=> {firstRespondTime=1}	0.7096	1.0000	1.2593
## 11	{firstRespondTime=1}	=> {lastPacketTime=1}	0.7096	0.8936	1.2593
## 12	{lastPacketTime=1}	=> {lastAckTime=1}	0.7096	1.0000	1.0844
## 13	{firstRespondTime=1}	=> {lastAckTime=1}	0.7760	0.9772	1.0597
## 14	{lastAckTime=1}	=> {firstRespondTime=1}	0.7760	0.8415	1.0597
## 15	{downAvBand=7, lastPacketTime=1}	=> {firstRespondTime=1}	0.1234	1.0000	1.2593
## 16	{downAvBand=7, firstRespondTime=1}	=> {lastPacketTime=1}	0.1234	0.8674	1.2224
## 17	{downAvBand=7, lastPacketTime=1}	=> {lastAckTime=1}	0.1234	1.0000	1.0844
## 18	{downAvBand=7, firstRespondTime=1}	=> {lastAckTime=1}	0.1386	0.9742	1.0564
## 19	{downAvBand=7, lastAckTime=1}	=> {firstRespondTime=1}	0.1386	0.8600	1.0830
## 20	{upAvBand=5, lastPacketTime=1}	=> {firstRespondTime=1}	0.1133	1.0000	1.2593
## 21	{upAvBand=5, firstRespondTime=1}	=> {lastPacketTime=1}	0.1133	0.8217	1.1580
## 22	{upAvBand=5, lastPacketTime=1}	=> {lastAckTime=1}	0.1133	1.0000	1.0844

FIGURE 31 –

Conclusion

Pour conclure, avec \LaTeX on obtient un rendu impeccable mais il faut s'investir pour le prendre en main.

Références

- [1] R&D DÉPARTEMENT DE CMCC. *Interface Specification of China Mobile Signaling Monitoring System(LTE Signal Collection Gateway Part)*.
- [2] Jianhua DU, Shiwen LU, Fangfeng ZHANG *Research of KQI Development Methodology in SQM*,2008.
- [3] Luning ZHAO, Zhuo SUN, Wenbo WANG *Mobile streaming QoE index system and quantify*,2012.
- [4] Rui WANG, Fei SU, Zhengdong HAN, Zilong CAI. *The Recessive Problem Mining and Optimization Research of Voice Service Based on User Behaviors*. édition, 2013.
- [5] Lianjiang ZHU, Bingxian MA, Xuequan ZHAO. *Clustering validity analysis based on silhouette coefficient*, 2010.
- [6] Hongyan WANG, Daiwen WU. *Discussion on digging algorithm of correlation rule for numerical attribute*, 2012.